Multimedia Tools and Applications
https://doi.org/10.1007/s11042-019-08129-4

Multi-subspace supervised descent method for robust face alignment

Jianwen Lou[1] · Xiaoxu Cai[1] · Yiming Wang[1] · Hui Yu[1] · Shaun Canavan[2]

Received: 13 November 2018 / Revised: 20 June 2019 / Accepted: 13 August 2019
Published online: 04 September 2019

Abstract
Supervised Descent Method (SDM) is one of the leading cascaded regression approaches for face alignment with state-of-the-art performance and a solid theoretical basis. However, SDM is prone to local optima and likely averages conflicting descent directions. This makes SDM ineffective in covering a complex facial shape space due to large head poses and rich non-rigid face deformations. In this paper, a novel two-step framework called multi-subspace SDM (MS-SDM) is proposed to equip SDM with a stronger capability for dealing with unconstrained faces. The optimization space is first partitioned with regard to shape variations using k-means. The generated subspaces show semantic significance which highly correlates with head poses. Faces among a certain subspace also show compatible shape-appearance relationships. Then, Naive Bayes is applied to conduct robust subspace prediction by concerning about the relative proximity of each subspace to the sample. This guarantees that each sample can be allocated to the most appropriate subspace-specific regressor. The proposed method is validated on benchmark face datasets with a mobile facial tracking implementation.

Keywords Unconstrained face alignment · SDM · Subspace learning · Cascaded regression

1 Introduction

Face alignment aims to automatically localize fiducial facial points (or landmarks). It is a fundamental step for many facial analysis tasks, e.g. facial recognition [19, 20], face frontalization [21, 22], expression recognition [11, 31], and face attributes prediction [7, 25]. These tasks are essential to Human-System Interaction (HSI) applications including driver-car interaction, human-robot interaction and mobile applications.

✉ Hui Yu
hui.yu@port.ac.uk

[1] School of Creative Technologies, University of Portsmouth, Portsmouth PO1 2DJ, UK

[2] Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA

The field of face alignment has witnessed rapid progresses in recent years, especially with the application and development of cascaded regression methods [2, 6, 27, 38, 39]. This kind of methods typically learns a sequence of descent directions from image features that move an initial shape towards the ground truth iteratively. Among various cascaded regression approaches for face alignment, SDM [27] has risen as one of the most popular approaches due to its high efficiency and the state-of-the-art performance. The approach is also theoretically sound to some extent with rigorous explanation from the perspective of optimizing a non-linear problem with Newton's method.

However, SDM has two main drawbacks: 1) It highly relies on the initialization and is prone to local optima. SDM is derived from Newton's method which leads to a local optima. If the initialised shape is far away from the target shape, the algorithm is prone to a poor local optimum (see Fig. 1a for an example). 2) It is likely to learn conflicting descent directions during optimization. As the feature extraction function in face alignment is not easy to describe, a simple function $h(x) = x^{-1}$ is used to illustrate it. Suppose the aim is to seek the optimal x ($x_* = 3.5$) that makes $h(x) = 0.286$ from a range of initial x (x_0). According to SDM, a descent map r can be calculated to move x_0 towards x_* iteratively using the following equation:

$$x_k = x_{k-1} - r(h(x_{k-1}) - h(x_*)) \tag{1}$$

For $x_0 \in$ [1:0.2:6] (0.2 is the interval), all of them can be moved closer to x_* with $r = -7$. Nevertheless, if $x_0 < 0$, e.g. $x_0 = -1$, then it will become farther away from x_* with $r = -7$ (see Fig. 1b).

Actually, only if initial points are close to each other and also target at the same destination, then the compatible descent directions can be learned via SDM. However, this strong prerequisite is very difficult to meet in face alignment, since face images vary from head poses and facial expressions, which are supposed to have different shape-feature relationships. This also leads to another issue of SDM: the algorithm is derived on a weak assumption that the non-linear feature extraction function (e.g. SIFT [13] or [17]) is identical for all the face images. As stated in [28], the feature extraction function is parameterized not only by facial landmark locations, but also by the images such as faces with different head poses and different subjects.

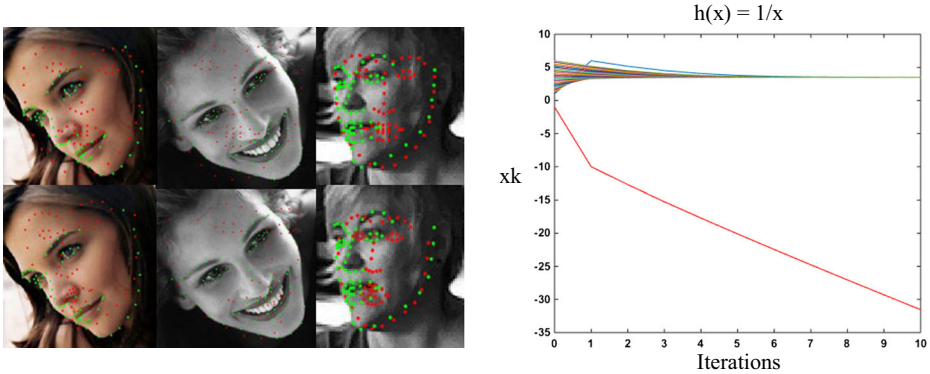


Fig. 1 **a** Failure cases of SDM due to poor initializations. Top row: initial shape, bottom row: results after four iterations. Red points: predicted landmarks, green points: ground-truth landmarks. **b** Initialization points that have conflicting descent directions

It can be inferred that one possible cause of above issues is that the face alignment task occupies multiple optimization subspaces, but these subspaces cannot be explained within a single optimization process. Although SDM has been extensively studied and further developed in the past few years, there are few works on this essential but relatively unexplored problem [8, 28, 29, 32, 35]. Xiong and De la Torre have made the same inference with this paper and proposed a global SDM (GSDM) [29] by domain partition in feature and shape PCA spaces for face tracking. However, that method is inappropriate for face alignment on still images as the decision of picking the suitable domain depends on ground-truth face shapes. The utilization of PCA also remains a big concern since it might result in un-estimated information loss. Recently, Zhang et al. [35] improves the GSDM by projecting both the feature and shape into a mutual sign-correlation subspace. Their method, however, has the same constraint as GSDM. Some other works resort to the multi-view approach – estimating head poses followed by face alignment on a particular view [12, 32]. The performance improves but the heuristic partition with respect to only head poses is still suboptimal because it neglects other shape deformations or appearance variations. Meanwhile, how to divide the pose range is a purely empirical step which often requires a lot of attempts.

To solve aforementioned problems, this paper proposes an efficient and novel alternative optimization subspace learning method – multi-subspace SDM (MS-SDM), which pushes SDM to the unconstrained face alignment application. The main contributions of our work are: 1) Discover optimization subspaces with a semantic meaning via applying an elegant unsupervised clustering algorithm – k-means on both shape and feature space. 2) Predict the subspace accurately by concerning about the relative proximity between the subspace and the sample. The proposed MS-SDM has been validated on challenging datasets which cover a wide range of head poses, facial expressions and facial appearances. Experimental results show the superiority of MS-SDM over SDM and GSDM.

2 Related work

A large number of works have been developed for face alignment which can be divided into two main categories: generative approaches and discriminative approaches.

Generative approaches, such as Active Appearance Models [4] and Constrained Local Models [5], first construct compact the shape and appearance spaces with Principal Component Analysis (PCA), then build a model instance to fit with the face image under a single optimization process. Although various improvements have been made, the drawbacks of this kind of approaches remain obviously: the expressive power of the built parameter space is limited and the final results heavily depend on the initialization.

Discriminative approaches don't build a parameter space beforehand, but alternatively they learn a direct mapping from image features to landmark locations [2, 27, 29, 38, 39]. Cascaded regression [2, 27, 38, 39] is a representative discriminative approach which has dominated the face alignment field in recent years due to its high efficiency and the state-of-the-art performance.

2.1 Face alignment with cascaded regression

Starting with a rough initial shape, cascaded regression predicts the shape increment from image features with a series of mapping functions, and update the shape iteratively. Cao et al.

[2] apply boosted ferns to learn both features and non-linear mappings which output promising results. In contrast, Xiong et al. [27] propose to use simple linear regression and hand-crafted features to accomplish cascaded regression which is named as Supervised Descent Method (SDM). Such simple configurations surprisingly generated state-of-the-art results. Recently, deep learning have also been applied on face alignment. The strong learning ability of deep models and the end-to-end learning mode enable deep learning based methods produce remarkable performance even for the most challenging datasets [15, 18, 30, 33, 34, 36]. However, deep learning methods always require a huge amount of training data and a very high computational capability, which make it difficult to be deployed on devices with limited resources. Ignoring on-going debates between deep learning and traditional methods, this paper makes a trade-off between efficiency and accuracy of the algorithm, based on the methods using SDM. Readers are referred to surveys [3, 23] for a comprehensive comparison of main-stream face alignment methods.

2.2 Face alignment with SDM based approaches

SDM produces the state-of-the-art performance with very elegant configurations, which has been regarded as an important benchmark method and triggers numerous new approaches in face alignment. As discussed above, only if the initializations are close to each other and the feature extraction function has a unique minimum, a sequence of generic descent directions can be learned via SDM. However, these prerequisite does not hold for faces under unconstrained conditions.

In [38], Zhu et al. starts each iteration by exploring a shape space rather than locking itself on a single initialization. This relaxes the optimization process from being affected by poor initializations to some extent and can lead to more robust face alignment. Nevertheless, the expressive power of a single regression in each iteration still remains a big concern. A few studies [12, 32] adopt intuitive multi-view approach to cover a wider optimization space and achieve a good performance. However, defining the optimization space according to head poses only is still sub-optimal since it neglects other shape deformations or appearance variations. In addition, the operation on dividing the head pose range is purely empirical and always needs a lot of attempts. Xiong et al. [29] theoretically analyzes this limitation of SDM and proposes Global SDM (GSDM) which partitions the optimization space into several domains based on reduced shape and feature. Although their method works well for face tracking and pose estimation, it is inappropriate for face alignment on still images as it requires the ground truth shape during prediction. Meanwhile, the reduced feature and shape space might lose some important information. To address the limitation of GSDM, Zhu et al. [39] proposes to learn a composition from predicted domain-specific shapes. This method performs well for faces with large poses and extreme expressions. Some other works resort to three-dimensional (3D) face modelling [8, 9, 26, 40] which requires additional 3D annotations of the training data. This paper presents an efficient alternative for optimization subspace learning that doesn't require any additional assumptions.

3 Methodology

In this section, the SDM method is recalled first and its limitations are theoretically analysed. Then, the proposed MS-SDM is introduced.

3.1 Supervised descent method

SDM converts the face alignment task which is originally a non-linear least squares problem into a simple least squares problem. It avoids computing Jacobian and Hessian with some supervised settings which significantly reduces the algorithm's complexity but at the same time generates state-of-the-art performance. Specifically, given a face image I and initial facial landmarks' coordinates $\mathbf{x_0}$, face alignment can be framed as minimizing the following function over $\Delta\mathbf{x}$:

$$f(\mathbf{x}_0 + \Delta\mathbf{x}) = \|h(\mathbf{x}_0 + \Delta\mathbf{x}, I) - h(\mathbf{x}^*, I)\|_2^2 \tag{2}$$

where $h(\mathbf{x}, I)$ represents the SIFT features (or HOG features) around the landmark locations $\mathbf{x}$ of image I. $\mathbf{x}_*$ represents the ground-truth landmark locations. Following Newton's method, with a second-order Taylor expansion, (2) can be transformed as:

$$f(\mathbf{x}_0 + \Delta\mathbf{x}) \approx f(\mathbf{x}_0) + \mathbf{J}_f(\mathbf{x}_0)^T \Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^T \mathbf{H}_f(\mathbf{x}_0)\Delta\mathbf{x} \tag{3}$$

where $\mathbf{J}_f(\mathbf{x_0})$ and $\mathbf{H}_f(\mathbf{x_0})$ are the Jacobian and Hessian matrices of f evaluated at $\mathbf{x_0}$. Differentiating (3) with respect to $\Delta\mathbf{x}$ and setting it to zero, the following equations can be obtained:

$$\begin{aligned} \Delta\mathbf{x} &= -\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_f(\mathbf{x}_0) \\ &= -2\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_h^{\mathbf{T}}(\mathbf{x}_0)(h(\mathbf{x}_0, I) - h(\mathbf{x}^*, I)) \\ &= -2\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_h^{\mathbf{T}}(\mathbf{x}_0)h(\mathbf{x}_0, I) + 2\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_h^{\mathbf{T}}(\mathbf{x}_0)h(\mathbf{x}^*, I) \end{aligned} \tag{4}$$

According to (4), the computation of the descent direction $\Delta\mathbf{x}$ requires $h(\mathbf{x}, I)$ to be twice differentiable or numerical approximations of the Jacobian and Hessian could be calculated. However, these requirements are difficult to meet in practice: 1) SIFT or HOG features are non-differentiable image operators; 2) numerically estimating the Jacobian or the Hessian in Eq. 4 is computationally expensive since the dimension of the Hessian matrix can be large and calculating the inverse of Hessian matrix is with $O(p^3)$ time complexity and $O(p^2)$ space complexity, where p is the dimension of the parameters to estimate [28]. Alternatively, SDM uses an identical pair of $\mathbf{R}$ and $\mathbf{b}$ to represent all face images' $-2\mathbf{H}_f^{-1}\,\mathbf{JT}\,h$ and $-2\mathbf{H}_f^{-1}\,\mathbf{JT}\,hh(\mathbf{x}_*, I)$ which are named as the descent direction. $\mathbf{R}$ and $\mathbf{b}$ define a linear mapping between $\Delta\mathbf{x}$ and $h(\mathbf{x_0}, I)$, which can be learned from the training set by minimizing:

$$\sum_{i=1}^{N} \left\| \Delta\mathbf{x}_*^i - \mathbf{R}h\left(x_0^i, I_i\right) - \mathbf{b} \right\|_2^2 \tag{5}$$

where, N is the number of images in the training set and $\Delta\boldsymbol{x}_*^i = \boldsymbol{x}_*^i - \boldsymbol{x}_0^i$. Since the ground-truth shape is difficult to be found in a single update step, a sequence of such descent directions denoted as $\{\mathbf{R}_k\}$ and $\{\mathbf{b}_k\}$ are learned during training. Then for a new face image, in each iteration k, the shape update can be calculated as:

$$\Delta\mathbf{x}_k = \mathbf{R}_k h(\mathbf{x}_{k-1}, I) + \mathbf{b}_k \tag{6}$$

The function $h(\mathbf{x}, I)$ is parameterized not only by $\mathbf{x}$ but also by face images [28], which highly depends on head poses, facial expressions, facial appearances and illuminations. Consequently, $\mathbf{R}$ and $\mathbf{b}$ may vary from different face images. Therefore, although SDM can generate

promising face alignment results in ordinary scenarios, they suffer from unconditional scenarios where faces have large head poses and extreme expressions.

In [29], the authors observe the same problem. They propose to partition the original optimization space into several domains based on reduced shape deviation $\Delta\mathbf{x}$ and feature deviation Δh. They prove that each domain contains a generic descent direction which can make the initial shape closer to the ground-truth shape for every sample belongs to it when both of the following conditions hold: 1) $h(\mathbf{x}, I)$ is strictly monotonic around $\mathbf{x}_*$ and 2) $h(\mathbf{x}, I)$ is locally Lipschitz continuous anchored at $\mathbf{x}_*$ with K ($K \geq 0$) as the Lipschitz constant**.** However, the solution proposed in [29] only satisfies the first condition above and is based on an assumption that $\Delta\mathbf{x}$ and Δh embedded in a lower dimensional manifold. Meanwhile, to predict the specific domain that a sample belongs to, the ground-truth shape $\mathbf{x}_*$ should be given. This is apparently infeasible during the testing stage as the ground-truth shape is actually what needs to be predicted.

3.2 Multi-subspace SDM

To address problems mentioned above, an alternative two-step framework – MS-SDM (see Fig. 2) is proposed. It first learns subspaces with semantic meanings from the original optimization space via k-means. Then, for each subspace, a particular linear regressor from face features to the shape update is learned. During testing, the sample will be assigned into the correct subspace with a pre-trained Naive Bayes classifier. It will then be allocated to a subspace specific regressor which gradually update the shape as:

$$\Delta\mathbf{x}_k = \mathbf{R}_{k,s} h(\mathbf{x}_{k-1}, I) + \mathbf{b}_{k,s} \tag{7}$$

where s represents the subspace label.

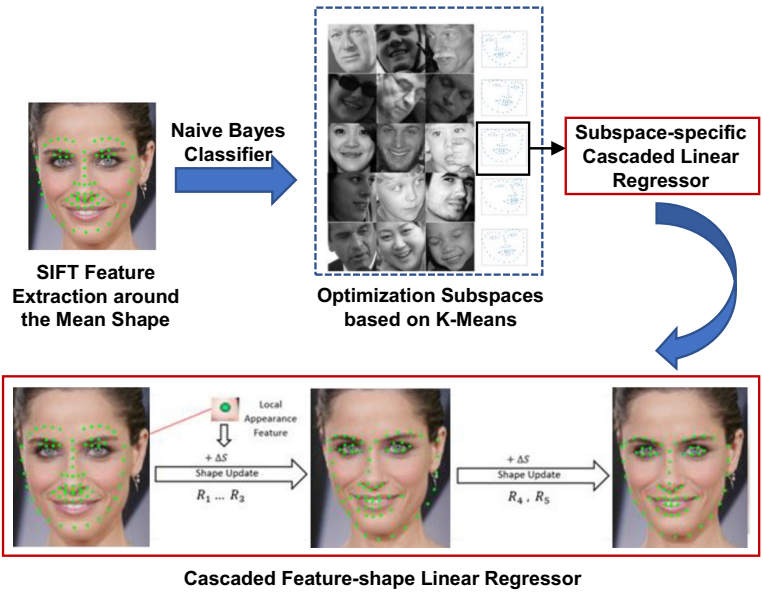


Fig. 2 The work pipeline of MS-SDM

3.2.1 Semantic subspace learning via K-means

To learn better optimization subspaces, samples which have the similar regression target $\Delta\mathbf{x}$ are assumed to fall inside the same optimization space and have compatible descent directions. Then, the classic clustering algorithm - k-means is applied on all training samples' $\Delta\mathbf{x}$ to automatically find out the key facial shape variations and divide the original training set into several subsets. In order to preserve all the useful information hidden in the shape space, the initial $\Delta\mathbf{x}$ of each sample is utilised during the clustering process. As shown in Fig. 3a, subsets generated in this way show quite high correlation with head poses. It can also be observed that each subset relates to a particular kind of head pose, such as left-profile face, right-profile face, left-rolling face and right-rolling face.

Since the face shape update $\Delta\mathbf{x}$ are predicted from the feature deviation Δh, the descent direction pair of $\mathbf{R}$ and $\mathbf{b}$ also describes the hidden relationship between $\Delta\mathbf{x}$ and Δh. Inspired by this intuition, k-means is further applied on Δh to find the feature-based optimization space partition. Surprisingly, the generated subspaces are highly consistent with the subspaces obtained from the head pose's point of view. The relevant results are shown in Fig. 3b. It indicates that samples in each subspace have close shape-feature relationships which are supposed to share a unified descent direction.

3.2.2 Robust subspace prediction with naive Bayes

As the aforementioned subspace learning relies on the ground-truth shape which will be unavailable during testing, the main difficulty of the final shape prediction arises as the prediction of the subspace that a sample belongs to. A straightforward solution to this problem is a multi-class classifier (e.g. Random Forest, SVM or Naive Bayes), which learns the class label from face appearance features.

In the test phase, a mean-face is placed onto the given face bounding box and SIFT features are extracted around each landmark (see Fig. 2). The concatenation of all extracted features are regarded as the appearance feature for subsequent classification. Random Forest was first

a Subspaces learned from Δx **b Subspaces learned from Δ*h***

Fig. 3 Comparison between learned subspaces from $\Delta\mathbf{x}$ and Δh. Each row represents a subset which contains three example images and the mean shape of all the samples in the subset. The cluster's amount of k-means is set as 5.

tested in our experiment due to its high performance in similar tasks. However, with this approach, a few samples were assigned inaccurately with a completely incompatible subspace, such as a left-profile face was assigned with a right-profile view regressor, which severely ruins the overall prediction accuracy.

The core reason behind this phenomenon is that Random Forest regards different subspaces equally. In particular, during training, it assigns the same loss punishment for any other sub-optimal subspace prediction. However, some sub-optimal subspace provides relatively similar initial-shape-indexed features and can predict similar shapes as the optimal one, which should be punished lighter. Therefore, a classification algorithm fits with this task should be able to identify the relative proximity between the sample and the subspace.

Naive Bayes appears to be a good option to this problem. A Naive Bayes classifier is the function that assigns a class label $y = C_k$ for some k as follows:

$$y = \arg\ \max_{k\in\{1,\ldots,K\}} p(C_k)\prod p(x_i|C_k) \tag{8}$$

where $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ represents the feature vector of a sample; $p(C_k)$ is a priori probability of class C_k, and $p(x_i|C_k)$ is the a posteriori probability of class C_k given the value of x_i. As Naive Bayes classifier assumes each feature x_i which is conditionally independent of every other feature x_j $(j \neq i)$, $p(\boldsymbol{x}|C_k)$ is equal to the product of all $p(x_i|C_k)$. The parameter $p(\boldsymbol{x}|C_k)$ can be regarded as the distance between the current sample to the class centre. If the sample is far away from the class centre, then $p(\boldsymbol{x}|C_k)$ is small, otherwise, $p(\boldsymbol{x}|C_k)$ turns large. Since $p(\boldsymbol{x}|C_k)$ directly contributes to the optimization process, the relative proximity between the sample and the class is then naturally embedded in the Naive Bayes Classifier. This can avoid assigning a sample with an incompatible subspace.

4 Experiments

Dataset Evaluations are performed on a widely applied benchmark dataset – 300 W [16] and NTHU Drowsy Driver Detection (NTHU-DDD) video dataset [24]. The dataset 300 W is a mixture of several well-known benchmark datasets, including AFW [37], LFPW [1], HELEN [10] and XM2VTS [14], which is challenging due to its images covering a very wide range of head pose, facial expression, appearance, occlusion and illumination. It unifies all the annotations with the 68-point mark-up and offers another challenging 135-image dataset named IBUG.

During the experiment, all the training samples from LFPW, HELEN and the whole AFW form the training set which has 3148 images in total. The testing set comprises of a common testing set and a challenging testing set, which has 689 images in total. The common testing set is composed of testing samples from LFPW and HELEN which have near-frontal head poses. IBUG is regarded as a challenging set as it is generally consisted of samples with large head poses and extreme facial expressions. Since the face detector's influence on the final face alignment results is not considered in this paper, the prescribed face bounding boxes provided by 300 W are used.

Evaluation metric The prediction error is measured as the average point-to-point Euclidean error normalised by the inter- pupil distance (the Euclidean distance between eyes' centres). For simplicity, the '%' is omitted.

Implementation During training, similar data augmentation as in [27] is applied to enlarge the training data and improve the model's generalization capability: the face bounding box of each training sample is randomly translated and scaled ten times. As samples in each subspace relate closely to a specific head pose, the mean shape of each subspace is calculated. Before prediction, each sample will be allocated a subspace-specific mean shape which is closer to the ground truth shape than the general mean shape. For subspace learning, the amount of clusters is altered from 3 to 8 and calculated the related error. The setting of 5 subspaces is shown to generate best results.

During the training process of the subspace classifier, it has shown that features indexed on multiple initial shapes can output higher prediction accuracy in comparison with features indexed on a single initial mean shape. This is probably due to that multiple initial shapes, which cover more points on the face region, can generate a larger feature pool and offer more information to the classifier. Therefore, shape-indexed features using all the subspace-specific mean shapes are extracted to train the subspace classifier.

4.1 Comparison with SDM

The released model of SDM was trained on private datasets and the training data has shown to be an important factor to the final performance of the model. What's more, there is no off-the-shelf GSDM model released. To enable fair comparison on the same benchmark dataset, we re-implement SDM and GSDM by ourselves. Our implementation achieves detection accuracy close to similar implementations that have been reported in some state-of-art works [34].

As shown in Table 1, the proposed MS-SDM outperforms SDM on all testing sets, especially on the challenging set. The challenging set contains many samples with large head pose and extreme facial expressions which have conflicting descent directions with near-frontal faces. As SDM can only learn an average descent direction which is prone to the descent direction shared by major samples (near-frontal faces), the learned descent direction cannot handle minor challenging samples. While MS-SDM classifies each sample into a subspace where samples share similar descent directions which guarantees even the challenging sample can get an effective descent direction. Figure 4 presents some example results which intuitively show MS-SDM's superiority over SDM.

4.2 Comparison with GSDM

GSDM offers an optimization space partition strategy for SDM which has demonstrated its effectiveness in real-time face tracking. To compare MS-SDM with GSDM, it is assumed that all the ground-truth shapes are known to make GSDM work even on still images. For both approaches, the subspaces are learned from the training set. Each subspace will be trained with a specific linear regressor. For fair comparison, the optimization space is partitioned into eight subspaces which are the same as that reported in [29]. As shown in Table 1, MS-SDM shows

Table 1 Comparison with SDM and GSDM

	Common Set	Challenging Set	Full Set
SDM	5.59	15.38	7.51
GSDM	5.39	12.57	6.80
MS-SDM	5.30	12.29	6.47

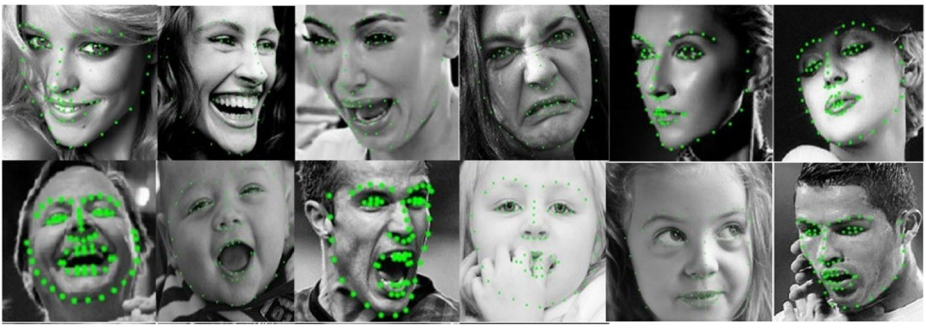


Fig. 4 Example results from the testing set

higher detection accuracy than GSDM on both testing sets. What's more, it learned subspaces without knowing ground-truth shapes which GSDM requires.

4.3 Tracking results on driver dataset

Figure 5 shows tracking results of our method on NTHU-DDD video dataset [24]. Detected facial landmarks can favour driver drowsiness detection which can further be used for facial analysis of drivers to reduce car accidents.

4.4 Facial Mobile tracking implementation

Based on MS-SDM, an Android facial tracking application was developed to track the user's face with 66 landmarks in real-time. The application can robustly track the face within a large range of head poses and facial expressions (see Fig. 6), while having low hardware requirements to run smoothly on an Android smart phone. It can also benefit many other useful mobile applications such as automated face makeup, personalised emoji generation and objective facial functionality assessment.

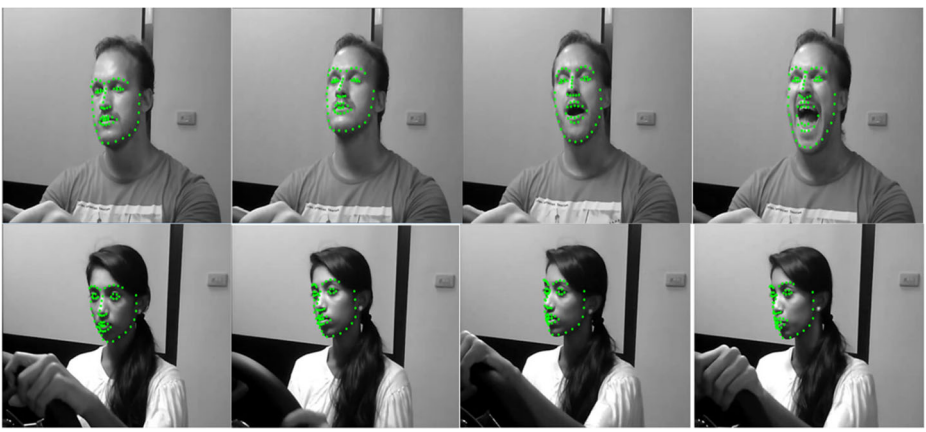


Fig. 5 Tracking results on NTHU Drowsy Driver Detection (NTHU-DDD) video dataset [24]

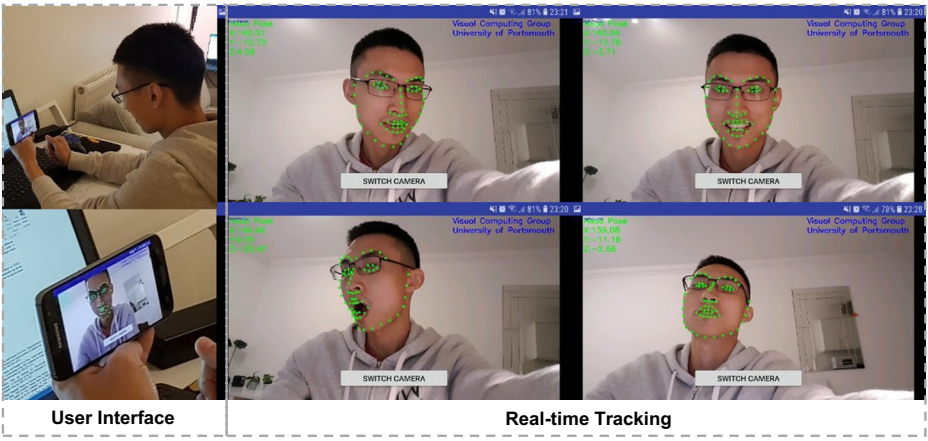


Fig. 6 Screenshots of the facial tracking mobile application based on MS-SDM

5 Conclusion

With a quite elegant formulation, SDM shows the state-of-the-art performance for face alignment under relatively controlled scenarios. As SDM is a local algorithm and prone to learn conflicting descent directions during training, it suffers from face images captured under unconstrained scenarios, where faces have large poses and extreme facial expressions. This paper proposes a novel two-step framework – MS-SDM which pushed SDM closer to unconstrained face alignment. Via applying k-means on the shape variations, semantic subspaces which have intuitive correlation with head poses are found. Then, using Naive Bayes classifier, each sample can be allocated the most suitable subspace-specific regressor. The proposed approach is validated on challenging datasets and a mobile facial tracking application. In future, we will apply deep learning techniques to extract more informative facial features or partition the feature-shape relationship into subspaces with clearer semantic meaning.

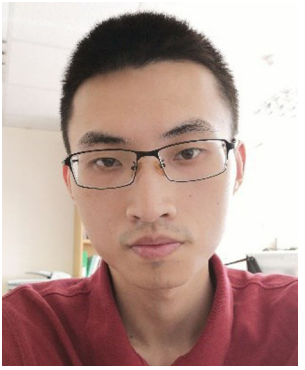
Acknowledgments This work was supported by the EPSRC through project 4D Facial Sensing and Modelling (EP/N025849/1), UoP RIDF2017 fund, the Emteq (https://emteq.net/) and was in part supported by the Open Fund of the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences (Y6S9011F51).

References

1. Belhumeur PN, Jacobs DW, Kriegman DJ, Kumar N (2013) Localizing parts of faces using a consensus of exemplars. IEEE Trans Pattern Anal Mach Intell 35(12):2930–2940
2. Cao X, Wei Y, Wen F, Sun J (2014) Face alignment by explicit shape regression. Int J Comput Vis 107(2):177–190
3. Chrysos GG, Antonakos E, Snape P, Asthana A, Zafeiriou S (2018) A comprehensive performance evaluation of deformable face tracking "in-the-wild". Int J Comput Vis 126(2–4):198–232
4. Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. IEEE Trans Pattern Anal Mach Intell (6):681–685

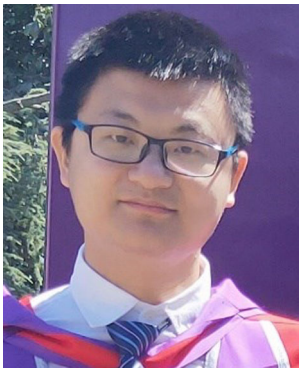
5. Cristinacce D, Cootes TF (2006) Feature detection and tracking with constrained local models. In Bmvc, Vol 1, No 2, p 3.
6. Guo S, Tan G, Pan H, Chen L, Gao C (2017) Face alignment under occlusion based on local and global feature regression. Multimed Tools Appl 76(6):8677–8694
7. Jian M, Lam KM (2015) Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition. IEEE Trans Circuits Syst Video Technol 25(11):1761–1772
8. Jourabloo A, Liu X (2015) Pose-invariant 3D face alignment. In: IEEE international conference on computer vision (ICCV), pp 3694–3702
9. Jourabloo A, Liu X (2016) Large-pose face alignment via CNN-based dense 3D model fitting. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 4188–4196
10. Le V, Brandt J, Lin Z, Bourdev L, Huang TS (2012) Interactive facial feature localization. In: European conference on computer vision. Springer, pp 679–692
11. Lian Z, Li Y, Tao J, Huang J, Niu M (2019) Expression Analysis Based on Face Regions in Read-world Conditions. Int J Autom Comput, pp 1–12
12. Liu Q, Deng J, Tao D (2016) Dual sparse constrained cascade regression for robust face alignment. IEEE Trans Image Process 25(2):700–712
13. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
14. Messer K, Matas J, Kittler J, Luettin J, Maitre G (1999) XM2VTSDB: The extended M2VTS database. In Second international conference on audio and video-based biometric person authentication 964:965–966
15. Saeed A, Al-Hamadi A, Neumann H (2018) Facial point localization via neural networks in a cascade regression framework. Multimed Tools Appl 77(2):2261–2283
16. Sagonas C, Tzimiropoulos G, Zafeiriou S, Pantic M (2013) 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: IEEE international conference on computer vision workshops (ICCV workshop), pp 397–403
17. Semwal VB, Mondal K, Nandi GC (2017) Robust and accurate feature selection for humanoid push recovery and classification: deep learning approach. Neural Comput & Applic 28(3):565–574
18. Shao X, Xing J, Lv JJ, Xiao C, Liu P, Feng Y, Cheng C, Si F (2017) Unconstrained Face Alignment Without Face Detection. In: IEEE Conference on computer vision and pattern recognition workshops (CVPR workshop), pp 2069–2077
19. Tao D, Guo Y, Li Y, Gao X (2017) Tensor rank preserving discriminant analysis for facial recognition. IEEE Trans Image Process 27(1):325–334
20. Tao D, Guo Y, Yu B, Pang J, Yu Z (2017) Deep multi-view feature learning for person re-identification. IEEE Trans Circuits Syst Video Technol 28(10):2657–2666
21. Wang Y, Yu H, Dong J, Stevens B, Liu H (2016). Facial expression-aware face frontalization. In Asian conference on computer vision. Springer, pp 375–388
22. Wang Y, Yu H, Dong J, Jian M, Liu H (2017) Cascade support vector regression-based facial expression-aware face frontalization. In: IEEE International Conference on Image Processing (ICIP), pp 2831–2835
23. Wang N, Gao X, Tao D, Yang H, Li X (2018) Facial feature point detection: a comprehensive survey. Neurocomputing 275:50–65
24. Weng CH, Lai YH, Lai SH (2016) Driver drowsiness detection via a hierarchical temporal deep belief network. In: Asian conference on computer vision. Springer, pp 117–133
25. Xia Y, Lou J, Dong J, Li G, Yu H (2018) SDM-based means of gradient for eye center localization. In IEEE International Conference on Pervasive Intelligence and Computing (PiCom), pp. 862–867
26. Xiao S, Li J, Chen Y, Wang Z, Feng J, Yan S, Kassim AA (2017) 3D-Assisted Coarse-to-Fine Extreme-Pose Facial Landmark Detection. In: IEEE Conference on computer vision and pattern recognition workshops (CVPR workshop), pp 2060–2068
27. Xiong X, De la Torre F (2013) Supervised descent method and its applications to face alignment. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 532–539
28. Xiong X, De la Torre F (2014) Supervised descent method for solving nonlinear least squares problems in computer vision. arXiv preprint arXiv:1405.0601
29. Xiong X, De la Torre F (2015) Global supervised descent method. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2664–2673
30. Yang J, Liu Q, Zhang K (2017). Stacked hourglass network for robust facial landmark localisation. In: IEEE Conference on computer vision and pattern recognition workshops (CVPR workshop), pp 2025–2033
31. Yu H, Liu H (2014) Regression-based facial expression optimization. IEEE Trans Hum Mach Syst 44(3): 386–394
32. Yu X, Lin ZL, Zhang S, Metaxas DN (2016). Nonlinear hierarchical part-based regression for unconstrained face alignment. In IJCAI, pp 2711–2717
33. Zhang J, Shan S, Kan M, Chen X (2014) Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In: European conference on computer vision. Springer, pp 1–16

34. Zhang Z, Luo P, Loy CC, Tang X (2014) Facial landmark detection by deep multi-task learning. In: European conference on computer vision Springer, pp 94–108
35. Zhang Y, Liu S, Yang X, Shi D, Zhang JJ (2016) Sign-correlation partition based on global supervised descent method for face alignment. In: Asian conference on computer vision. Springer, pp 281–295
36. Zhao Y., Tang F, Dong W, Huang F, Zhang X (2018) Joint face alignment and segmentation via deep multi-task learning. Multimed Tools Appl 1–18
37. Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 2879–2886
38. Zhu S, Li C, Loy CC, Tang X (2015) Face alignment by coarse-to-fine shape searching. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 4998–5006
39. Zhu S, Li C, Loy CC, Tang X (2016) Unconstrained face alignment via cascaded compositional learning. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 3409–3417
40. Zhu X, Lei Z, Liu X, Shi H, Li SZ (2016) Face alignment across large poses: A 3d solution. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 146–155



Jianwen Lou received his M.Sc. degrees from Ocean University of China in 2016. He is currently pursuing the Ph.D. degree in the School of Creative Technologies in University of Portsmouth. His research interests include 2D facial tracking, facial animation and machine learning.

Xiaoxu Cai received her M.Sc. degrees from Ocean University of China in 2016. She is currently pursuing the Ph.D. degree in the School of Creative Technologies in University of Portsmouth. Her research interests include 3d face reconstruction, face recognition and deep learning.



Yiming Wang is a PhD student in the School of Creative Technologies at the University of Portsmouth. His research interests include machine/deep learning and automatic facial expression analysis. He won the best paper prize at the International Conference on Human System Interaction (HSI 2015).

Hui Yu is a Professor with the University of Portsmouth, UK. His research interests include vision, computer graphics and application of machine learning to above areas, particularly in human machine interaction, image processing and recognition, Virtual/Augmented reality, 3D reconstruction, robotics and geometric processing of human/facial performances. He is Associate Editor of IEEE Transactions on Human-Machine Systems and the Neurocomputing journal.

Shaun Canavan received his PhD in Computer Science from Binghamton University in 2015. During the summer of 2012, he was a visiting faculty member of the Air Force Research Lab in Rome, New York where he worked on 3D object reconstruction from 2D images. After his PhD, he was co-director of the Graphics and Image Computing Lab, as well as a Research Assistant Professor in the Freshman Research Immersion program at Binghamton University where he mentored undergraduates on research in biometrics, HCI, and machine learning. Canavan has published in top conferences such as CVPR, FG, and BTAS. He joined the Department of Computer Science and Engineering at the University of South Florida (USF) in Fall 2017.