2017 IEEE 12th International Conference on Automatic Face & Gesture Recognition

Constrained Ensemble Initialization for Facial Landmark Tracking in Video

Christy (Yuan) Li, Tadas Baltrušaitis, Louis-Philippe Morency
Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA, USA

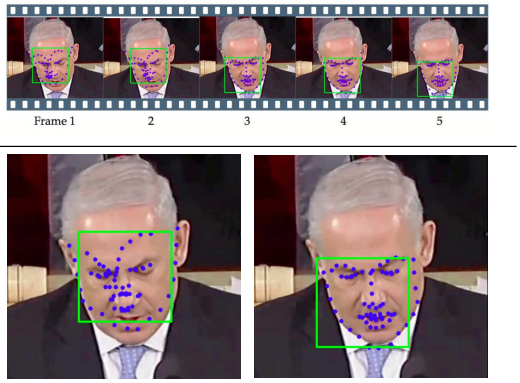
***Abstract*— Accurate and robust facial landmark tracking is a crucial step for face recognition and affect analysis systems. We often want to not only detect facial landmarks in images but to be able to track them reliably and consistently over time. Recently there has been an increase in research interest in facial landmark detection, especially in cascaded regression based methods such as the Supervised Descent Method (SDM). However, while facial landmark detection in images has improved significantly, comparably very little attention has been given to the task of landmark detection/tracking in videos. In our work we present a novel initialization procedure that can help with cascaded regression based facial landmark detection and tracking. Our initialization technique exploits the fact that cascaded regression is sensitive to initialization noise, especially in the presence of out-of-plane head pose variation, e.g. when a person is looking down when reading or during fast head motion. Our approach allows to learn good candidates for initialization, that we exploit in our tracking framework. We evaluate our technique on 300VW dataset – a large publicly available corpus of *in-the-wild* videos and demonstrate its effectiveness for a number of cascaded-regression landmark detection approaches.**

I. INTRODUCTION

Accurate and robust facial landmark tracking is critical in many facial analysis applications [4]. This includes facial behavior analysis, human-computer and human-robot interaction, affective computing, lip reading, and surveillance. Many of these applications also require not only to detect landmarks in images but also track these landmarks over time in videos.

While facial landmark detection has improved significantly in the past years, comparably very little attention has been given to the task of landmark detection/tracking in videos [8]. This is especially true for difficult to track *in-the-wild* scenarios that contain large head motion, occlusion, and changes in illumination.

While cascaded regression [26], [9], [29] has been very popular for image-based landmark detection due to its accuracy, the predicted landmarks are often unstable and the performance depends a lot on initialization (see Figure 1), meaning that even a small perturbation in initial landmark locations lead to largely different convergence. Furthermore, this often results in "jittery" tracking in videos.



frame 2 with different initial bounding boxes

Fig. 1: Visualization of initial bounding boxes and landmark predictions in 5 frames of a testing video. The second row shows the second frame with different initial bounding boxes in larger scale.

In our work we present a novel initialization procedure that can help with regression based facial landmark detection and tracking. Our approach enables smart initialization of the model based on landmark detections in the previous frame, with a linear scaling of time complexity. We demonstrate the benefit of our approach on 300VW [18] – a large publicly available dataset of *in-the-wild* videos. Furthermore, we demonstrate how our approach can be used to initialize a range of modern landmark detection methods, with improved results for all of them.

The paper is structured as follows: we first present the relevant previous work (Section II); in Section III we describe the cascaded regression model we use; this is followed by a discussion of our proposed initialization strategy in Section IV; we then describe our experimental procedure and results in Sections V and VI.

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA 2014-14071600011. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

978-1-5090-4023-0/17 $31.00 © 2017 IEEE
DOI 10.1109/FG.2017.88

II. RELATED WORK

Most of the previous work on facial landmark detection has concentrated on detecting landmarks in still images [16], [15], [24]. We review the work done in this domain by first reviewing the currently very popular cascaded regression methods. We continue by discussing the model (re)initialization strategies used to make the facial landmark detections more robust, and finally we survey the methods that have been proposed to track facial landmarks in videos.

A. Cascaded regression

Arguably, the most popular method for facial landmark detection at the moment is the cascaded regression framework [9], [26], [21]. In this framework facial landmark detection is updated in a cascaded fashion. That is, the landmark detection is continually improved by applying a regressor on appearance given the current landmark estimate as performed by Cao et al. in explicit shape regression [7]. Other cascaded regression approaches include the Supervised Descent Method (SDM) [26] which uses SIFT [13] features with linear regression to compute the shape update. Coarse-to-Fine Shape Searching (CFSS) [29] model attempts to avoid a local optima in cascaded regression by performing a coarse to fine shape search. Project out Cascaded regression (PO-CR) [21] presents an alternative formulation of cascaded regression in the Projected Out space and updates the shape model parameters rather than predicting landmark locations directly. Robust Cascaded Pose Regression (RCPR) [6] was proposed to improve the accuracy of cascaded regression to outliers by detecting occlusions explicitly and using robust shape-indexed features.

Recent work has also used deep learning techniques in a cascaded regression framework to both extract the visual features and to perform the model parameter update. Coarse-to-Fine Auto-encoder Networks (CFAN) [28] use visual features extracted by an auto-encoder together with linear regression. Sun et al. [19] proposed a Convolutional Neural Network (CNN) based cascaded regression approach for sparse landmark detection, however while their approach is robust it is not very accurate. More recently a Mnemonic Descent Method has been proposed, which performs cascaded regression using a CNN to extract visual features and a Recurrent Neural Network to perform the parameter update[20].

Our work builds on top of these advances in cascaded regression and presents a way of making them more accurate in case of landmark detection in difficult videos.

B. Reinitialization

While very accurate, cascaded regression is often not very stable. Slight differences in initialization may lead to drastically different predicted landmarks (see Figure 1). Several strategies have been proposed to deal with this issue . Cao et al. [7] and Dollár et al. [9] propose to use a set of different initializations to perform the inference and take the median of their outputs as the final prediction. An extension to this, called smart restarts, has been proposed by Burgos-Artizzu et al.[6] where the cascade is restarted if it starts drifting from the consensus just after a few iterations. While the above approaches perform the same initialization strategy irrespective of an image or expected location of the landmarks our work proposes smart sampling based on face orientation.

A number of alternative approaches have been proposed to use initially detected head pose [27], facial landmarks [28] or both [10] to reduce the search space for following landmark detection. However, all of these approaches have been designed with static images in mind and do not consider the temporal consistency required for tracking landmarks in videos. Applying them could lead to jittery and non smooth tracking.

C. Video landmark detection

The majority of work on facial landmark detection so far has concentrated on images. Tracking landmarks in videos, especially *in-the-wild*, has been less explored and has mostly concentrate on qualitative assessments on short videos due to a lack of benchmark datasets [8]. However, with the recent availability of the 300 Videos in the Wild (300VW) dataset [18], there have been a number of methods evaluated for the task of facial landmark tracking under difficult scenarios.

Chrysos et al. [8] evaluate a number of modern facial landmark detection methods in tracking pipelines based on face detection, model free tracking, and hybrid methods. They find that model free tracking (tracking of initially detected landmarks). rather than re-detecting landmarks each frame, leads to similar performance, with hybrid approaches leading to only a marginal increase in performance. However, they find that model free trackers do not show good performance in very difficult scenarios including occlusions, and large pose and illumination variation.

Majority of facial landmark tracking methods fall into one of two categories [8]: face detection in every frame followed by landmark detection; face detection in first frame followed by landmark detection in successive frames using the fitting result from previous frame (with potential re-detection upon failure) [2], [3]. Majority of approaches use the latter method that does not exploit the temporal nature of the video and does not enforce any temporal consistency. For example, Rajamanoharan et al. [14] use a multi-view Constrained Local Model and initialize it from previously detected landmarks. Such initialization technique is also used by Wu and Ji [25]. Uricar et al. [23] enforce the temporal consistency of face detection using a Kalman Filter, before applying fine grained

landmark detection, but they do not consider the consistency of finally detected facial landmarks. We take the landmark detection in videos one step further by introducing a method for robust landmark detection in particularly difficult videos.

III. Cascaded Regression

In this session, we describe Cascaded regression approaches to landmark detection. The problem of predicting facial landmark can be treated as updating a current shape iteratively from an initial shape towards ground truth landmarks. Cascaded regression treats each such iterative update as a regression problem and learns a mapping from current shape estimate and the corresponding appearance to ground truth shape. Xiong et al. [26] provides theoretical guarantees for the estimated shape moving closer to ground truth shape through iterations. During training cascaded regression learns a set of regressors – typically one for each iteration. During testing, it applies the regressors for shape estimation in a cascading manner until convergence or after reaching the maximum number of iterations.

Let $I = \{I_1, I_2, \ldots, I_N\}$ denote N training images, $S^* = \{S_1^*, S_2^*, \ldots, S_N^*\}$ denote the corresponding ground truth shapes, and $S^0 = \{S_1^0, S_2^0, \ldots, S_N^0\}$ denote the initial shapes. Cascaded regression learns a mapping f^k at iteration k ($k \in \{1, 2, \ldots, K\}$) such that the error between shape estimation S_i^k at the k_{th} iteration and ground truth shape S_i^* is minimized, where the i means the i_{th} image in training data. To estimate mapping given all training data, the least square error minimization problem is formulated as follows:

$$min \sum_{i=1}^{N} \|S_i^* - S_i^k - \Delta S_i^k\|^2 \tag{1}$$

where $\Delta S_i^k = f^k(S_i^k)$, f^k is a function that models relations between shape estimation at k_{th} iteration and ground truth shape. Usually, f involves extracting appearance features from an image at the current shape S_i^k and applying a linear operation on them. Note that in cascaded regression, each iteration learns a different f and updates the shape estimate at iteration k by applying f^k on it.

In testing, the shape estimation at iteration k is updated by adding the movement of landmarks approximated by f. The update function is as follows:

$$S_i^{k+1} = S_i^k + f(S_i^k) \tag{2}$$

In recent years there have been a number of cascaded regression methods proposed that show excellent performance in predicting facial landmarks in images [26], [7], [29], [22]. We will briefly discuss one of the most popular cascaded regression methods – Supervised Descent Method.

A. Supervised Descent Method (SDM)

In order to minimize the l_2 error between current and ground truth shapes in Equation 1, the elements need to be differentiable and the Hessian matrix and Jacobian matrix of the objective function have to be calculated [26]. This is often not analytically possible and/or very computationally expensive. SDM circumvents these difficulties by using a linear function to approximate relations between current shapes and ground truth shapes, without explicitly solving the least squares problem.

$$\Delta S_i^k = f(S_i^k) = R^k \phi^k(S_i^k) + b^k \tag{3}$$

Above, R^k term is the linear mapping that should be learned at iteration k so as to minimize the least square loss in Equation 1, b^k is a bias term that is also learned at iteration k, ϕ^k is the feature extraction function which takes the current shape estimate S_i^k as input and outputs the features extracted at the shape. Scale invariant feature transform (SIFT) [13] features are commonly used in cascaded regression, with different scales used at different iterations [26], [22].

During testing, the shape estimation at iteration k is updated by applying the linear mapping R^k and bias term b^k on it. The update function is then performed using Equation 2.

By directly learning a linear mapping from current shape estimations to ground truth shapes, SDM is computationally efficient. [26] shows that it leads to good performance on several applications such as facial landmark detection. However, as seen in Figure 1 the model performance is very sensitive to noise in model initialization.

IV. Ensemble initialization Cascaded Regression

Usually cascaded regression is performed from a single bounding box detection of a face, which is used to initialize the landmark locations based on an assumed *mean face*. However this sometimes leads to unstable performance as an initialization a couple of pixels away can lead to drastically different landmark detections. Ensemble fitting can help avoid such outliers by averaging a number of predictions from different initializations. In this section, we discuss two such ensemble learning algorithms, namely ensemble initialization and constrained ensemble initialization.

In regular cascade regression, given a video V with F frames, each frame's landmarks are initialized as x_0^i and updated by sequentially applying the update rule described in Equation 2 to arrive at the prediction result for each frame $X_p = \{x_p^i\}$. In ensemble setting, each frame is initialized M times. Let x_{0m}^i denote the m_{th} initialization of the i_{th} frame in video V. For the whole video, the initial shapes are denoted as $X_0 = \{x_{0m}^i\}$. After feeding X_0 into the cascaded regression model, we have $M \times F$ predictions, denoted

Algorithm 1 Select a shape based on mean shape

procedure MEANSHAPE
  shapeNum $\leftarrow$ # candidate shapes
  S $\leftarrow \{S_s\}$ where s = $\{1, \cdots,$ shapeNum$\}$
  $S_{Mean} \leftarrow$ mean of elements in S
  distance $\leftarrow \{\}$
  selectedShapes $\leftarrow \{\}$
  for s $\leftarrow \{0, \cdots,$ shapeNum$-1\}$ **do**
    distance $\leftarrow$ insert *getDistance(S_s, S_{Mean})*
  end for
  $d_{Mean} \leftarrow$ mean of distance
  $d_{Std} \leftarrow$ standard deviation of distance
  for s $\leftarrow \{0, \cdots,$ shapeNum$-1\}$ **do**
    if distance$[s] - d_{Mean} \leq \alpha * d_{Std}$ **then**
      selectedShapes $\leftarrow$ insert S_s
    end if
  end for
  finalShape $\leftarrow$ mean of selectedShapes
  return finalShape
end procedure
procedure GETDISTANCE
  a $\leftarrow$ *shape matrix*
  b $\leftarrow$ *shape matrix*
return Euclidean distance between a and b
end procedure

as $X_p = \{x^i_{pm}\}$. We call the M predictions of each frame as candidate predictions of that frame since they are used for computing a final prediction by the selection algorithms that we will illustrate next.

A. Ensemble Initialization

Ensemble initialization exploits the information given by multiple candidates of a certain frame without knowing the prediction of previous or consecutive frames. The algorithm uses the shape and votes for shapes that have mean or median Euclidean distance to the mean of provided shapes.

That is, given the predicted shapes of frame i, $X^i_p = \{x^i_{pm}\}$, Ensemble selection algorithm first computes the mean shape of all provided shapes, X^i_{mean}. Then a distance vector is formed by computing the Euclidean distance between every shape and the mean shape. Let $Dist = \{dist^i_{pm}\}$ denote the Euclidean distance between the m_{th} predicted shape and X^i_{mean} of the i_{th} frame. Ensemble selection algorithm chooses the shape that has either median $dist^i_p$ or mean $dist^i_p$ in $Dist$. To choose a shape that has mean $dist^i_p$ in $Dist$, one can first compute the mean and standard deviation of $Dist$, denoted as d_{mean}, d_{std}, and then select shapes whose $dist^i_p$ is within α times of standard deviation of $Dist$. Let $S^i_p = \{s^i_{pm}$ if $dist^i_{pm} - d_{mean} \leq \alpha \cdot d_{std}\}$, α is a hyper-parameter that can be optimized through validation. The final shape can then be computed by calculating the mean of shapes in S^i_p. The process is illustrated in detail in Algorithm 1.

B. Constrained Ensemble Initialization

In practice ensemble initialization leads to more accurate landmark detection. Theoretically, the more bounding boxes, the more candidates are provided for final inference, and thus the higher possibility to produce good result, assuming a good selection method exists. However, running too many predictions for a single frame leads to a high computational cost. This leads us to a desire to reduce the number of initializations for efficiency. Existing methods [30] find the number of initializations that balance computational efficiency and model accuracy. However, they treat the contribution of each initialization in the same way.

Experimentally, we observed that different bounding boxes, even with same distance from the ground truth, lead to very different predictions. This is especially true for non-frontal faces, where a good initial bounding box is crucial for an accurate prediction. An example of this is presented in Figure 1. Where, initializations below the ground truth, consistently outperform initializations above it. This is the case for majority of images with downwards head pose.

We further explore the relationship between the head pose, initial bounding box positions provided by a face detector, and cascade regression accuracy in an experiment. First, the detected bounding boxes are often biased based on the head pose of the person, for example if the person is looking towards right, the detected bounding box tends to be biased to the left, with the opposite effect for leftwards facing pose. The same effect also appears in upwards and downwards head poses. In our experiments, we find that shifting bounding boxes appropriately, allowed us to increase or decrease landmark detection accuracy.

We thus propose a method to select useful bounding boxes based on head pose. We manually fix the bounding box of certain locations based on head pose estimation. Let $P = \{p_1, p_2, \ldots, p_m\}$ denote m head pose categories, $B = \{b_1, b_2, \ldots, b_w\}$ denote available bounding boxes that are placed at different directions of a given initial bounding box, $r_i = p_j, j = \{1, 2, \ldots, m\}$ denote the head pose of the i_{th} frame in a video which falls in one of m pose categories. A set of bounding boxes $B(r_i) = \{b_{i1}, b_{i2}, \ldots, b_{iq}\}$ is then selected based on the estimated head pose in the previous frame.

We propose two methods to select bounding boxes based on head pose. One method is to manually match head pose with a set of selected bounding boxes based on hypothesis that the selected bounding box should be at the opposite direction of face direction. This method does not depend on training data and thus can be applied directly on various existing facial landmarks prediction algorithms.

However, different facial landmarks prediction algorithms may have a different pattern between head pose and expected bounding boxes. Thus, we further

propose an algorithm to learn such matching from training data. The algorithm first computes prediction from all available bounding boxes for every frame in a training video and then computes error of these predictions from their corresponding ground truths. Let $E \in R^{n \times w}$ denote the errors for n frames and w bounding boxes. We generate a ranking for each bounding box for each frame based on error values in ascending order. Then summarize for every bounding box, how many times it has been ranked as k for a certain head pose p_j, where k = $\{1, 2, \ldots, w\}$, j = $\{1, 2, \ldots, m\}$. Then, by using similar approach as top-k precision, we sort the bounding boxes in descending order of total number of times being ranked as top-k for every head pose category, and select fixed number of bounding boxes for every head pose category. This also allows us to control the number of initializations we want to perform allowing for a trade-off between computational cost and accuracy.

Since we are predicting landmarks in videos, we propose to use the prediction of previous frame to compute head pose and used as the estimated head pose for next frame. By using this temporal information in videos, constrained ensemble initialization can be applied without extra computational cost, making it a good fit for any facial landmark detection algorithms for videos. To estimate the head pose from detected landmarks we use an assumed mean face shape and solve the Perspective-n-Point problem.

Lastly, the constrained ensemble initialization can be run with a consistent time complexity as O(NM), where N is the number of initializations and M is the complexity of the landmark detector. In practice, the number of initializations in our constrained method does not exceed 15. Besides, the proposed method is universal across different state-of-the-art landmark detection methods and is able to improve the performance of recent SDM, PO-CR, and Chehra methods. We also believe that it could be applied on other recent landmark estimation methods.

V. Experiments

A. Dataset

This section describes the dataset we used for facial landmark tracking. 300VW [18] has 114 videos and training and testing split provided with the dataset. 50 videos are used for training. The rest of the videos are in the test set and are categorized in three scenarios. The first scenario includes videos in laboratory and well-lit conditions, with people displaying arbitrary expressions and head poses. The second scenario consists of videos of people recorded in unconstrained conditions such as varied illumination, dark rooms and overexposed shots. The third scenario focuses on completely unconstrained conditions including illumination and occlusions such as occlusions by hand.

For additional training data we used the popular 300W meta dataset [16]. 300W consists of four datasets, namely Helen [12], LFPW [5], AFW [31] and iBug [17]. Helen has 2330, LFPW 1024, AWL 337, and iBUG 135 images.

We used all four 300W datasets in addition to subsampled frames from video 300VW training dataset for our model training. For testing, we evaluated our model on all of the three categories of video testing set.

B. Face Detector

For tracking facial landmarks in videos we performed face detection in every frame using HOG-SVM face detector from the dlib library [11], which has been corrected to produce bounding boxes for the outline of the face encapsulating the 49 landmarks of interest. For images that HOG-SVM fails to detect a face region, we search for a bounding box in the nearest frames. The HOG-SVM face detection gives exactly one detection of bounding box for one frame.

C. Landmark Detectors

To explore our initialization strategies we used three modern facial landmark detection algorithms: **SDM** [26] implemented in the Intraface toolkit; our re-implementation of SDM based on the Menpo toolkit [1]; **Discriminative Response Map Fitting** (DRMF) implemented in the Chehra toolkit [2]; and **Project-out Cascaded Regression** (PO-CR) – a state-of-the-art cascaded regression approach [22]. Since these algorithms were trained without using 300VW dataset, we further train an SDM model by using the 300VW dataset.

D. SDM Landmark Detector

We trained and SDM model using the Menpo toolbox [1] on the full 300W dataset and subset of frames from 300VW video training set (we choose 1 frame from every 100). Since in videos, consecutive frames are very similar, using all of them could lead to overfitting. Validation shows that subsampling rate of 0.01 leads to good performance. For every training image, 10 perturbations of bounding boxes are generated by adding the noise distribution of our face detector. We assume that the noise can be modeled using a Gaussian on the of scale and translation errors between the detected bounding boxes and ground truth bounding boxes. Note that the bounding boxes generated in training is different from bounding boxes in testing in which ensemble or constrained ensemble initialization methods are used.

We trained a 6 iteration model that operates on 3 scales, with the scale staying the same for two iterations. The scales are in increasing order throughout training, which allows the model to check the landmark regions at increasingly larger scales and thus obtain more detailed information.

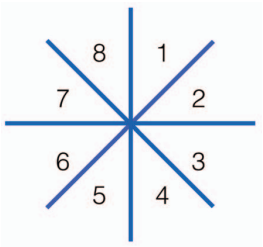


Fig. 2: Head pose categories. The head rotation space is divided into 8 categories using horizontal, vertical lines and lines with slope 1 and -1.

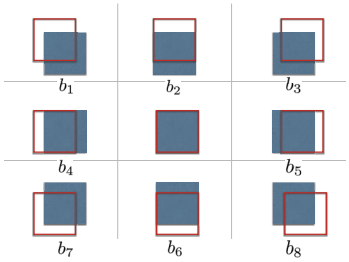


Fig. 3: 8 available bounding boxes for a certain scale noise. The bounding boxes are generated by shifting the same amount of noise from detected bounding box toward 8 directions: top-left, top, top-right, left, right, bottom-left, bottom, bottom-right.

E. Initialization experiments

We compare the initialization methods - single, ensemble, and constrained ensemble initialization.

For single initialization we use the initial bounding box obtained from the face detector. For ensemble initialization, we shifted the original bounding box by a fixed amount of noise in 8 directions: top-left, top, top-right, left, right, bottom-left, bottom and bottom-right. To account for scaling noise we also up-scale and downscale the shifted bounding boxes, leading to a total of 24 initial bounding boxes.

In the ensemble initialization experiment, the median position of each landmark from the predicted landmarks of 24 initial bounding box was computed and used as a final prediction for that landmark.

For constrained ensemble initialization, we find the pattern between face rotation and bounding boxes that produce good predictions. Specifically, we manually set the face rotation into 8 categories: looking towards top-left, top, top-right, left, right, bottom-left, bottom and bottom-right. Figure 2 shows the 8 rotation categories. We denote them as $\{r_1, r_2, \ldots, r_8\}$. Then for each rotation category, we fix a subset of bounding boxes for multiple initialization. We denote bounding boxes as in Figure 3. For one scale, there are $\{b_1, b_2, \ldots, b_8\}$ bounding boxes. In total, we use 3 scales, with positive and negative

rotation category	bounding boxes
1	$\{b_2, b_3, b_5, b_6, b_9\}$
2	$\{b_2, b_3, b_5, b_6, b_8, b_9\}$
3	$\{b_2, b_3, b_5, b_6, b_8, b_9\}$
4	$\{b_3, b_5, b_6, b_8, b_9\}$
5	$\{b_1, b_4, b_5, b_7, b_8\}$
6	$\{b_1, b_2, b_4, b_5, b_7, b_8\}$
7	$\{b_1, b_2, b_4, b_5, b_7, b_8\}$
8	$\{b_1, b_2, b_4, b_5, b_7\}$

TABLE I: The correspondence of head pose and bounding boxes used as initialization for ensemble initialization

rotation category	number of samples	bounding boxes
1	83	$\{b_2, b_3, b_5, b_6, b_9\}$
2	79	$\{b_2, b_3, b_5, b_8, b_9\}$
3	100	$\{b_2, b_3, b_5, b_8, b_9\}$
4	45	$\{b_3, b_5, b_6, b_8, b_9\}$
5	44	$\{b_1, b_4, b_5, b_7, b_8\}$
6	50	$\{b_1, b_2, b_4, b_5, b_7\}$
7	98	$\{b_1, b_2, b_4, b_5, b_7\}$
8	142	$\{b_1, b_4, b_6, b_7, b_8\}$

TABLE II: The correspondence of head pose and bounding boxes used as initialization for ensemble initialization through learning with fixed number of selected bounding boxes.

noise to the fixed 8 bounding boxes. So there are in total 24 bounding boxes $\{b_1, b_2, \ldots, b_{24}\}$ where $\{b_1, \ldots, b_8\}, \{b_9, \ldots, b_{16}\}, \{b_{17}, \ldots, b_{24}\}$ belong to no scale, positive scale, negative scale transformation respectively. The correspondence of face rotation and bounding boxes for no scale transformation used for predicting results are shown in table I. Bounding boxes of different scale noise go to the same labels as bounding boxes of no scale noise.

In the ensemble initialization, we manually select suitable bounding boxes for a certain face rotation based on hypothesis and observations from training data. Since training data is available, we also performed learning on training data to exactly discover the most informative bounding boxes.

First, by running the algorithm proposed in Section IV-B, we obtain the statistics of a number of top-3 rankings of each bounding box $b = \{b_1, b_2, \ldots, b_w\}$ for each frame in training data. By further choosing 5 bounding boxes from top-3 rankings, we obtain the matching of bounding boxes and head pose shown in Table II

VI. Results and Discussion

A. Ensemble Initialization

In this section we show the results of using a single initialization versus ensemble initialization. As an error metric we used the Area Under the Curve of the cumulative error curve on the 300VW test dataset. For easy comparison with other baseline facial landmark detection methods such as SDM and Project-out Cascaded Regression, we used the 49 point

	baseline	multiple initialization
Chehra	0.0522	**0.0538**
IntraFace	0.0479	**0.0500**
SDM*	0.0485	**0.0570**
PO-CR	0.0605	**0.0614**

TABLE III: AUC of Chehra, IntraFace, SDM and PO-CR using single initialization and ensemble initialization. SDM* means we used our implementation of SDM based on Menpo code [1].

	single initialization	ensemble initialization	constrained ensemble initialization
Chehra	0.0522	0.0538	**0.0558**
IntraFace	0.0479	0.0500	**0.0521**
SDM*	0.0485	**0.0570**	0.0540
PO-CR	0.0606	0.0614	**0.0620**

TABLE IV: AUC of Chehra, IntraFace, SDM and PO-CR using single initialization, ensemble initialization and constrained ensemble initialization. SDM* means we used our implementation of SDM based on Menpo code [1].

configuration for computing cumulative error curve of predicted landmarks. Table III shows the results of our experiment and it can be seen that for all of the methods, multiple initialization achieves larger AUC than the single initialization method.

B. Constrained Ensemble Initialization

Table IV shows the experiment results of using single initialization, ensemble initialization and constrained ensemble initialization. We can see from the table that Chehra, IntraFace and PO-CR have the highest prediction AUC when using our proposed constrained ensemble initialization, while SDM has a slightly larger AUC when using ensemble initialization.

C. Constrained Ensemble Initialization Through Learning

Table V shows the AUC of each method on four facial landmark prediction algorithms. We can see that Chehra, IntraFace and PO-CR achieve higher AUC when using learned initial bounding boxes. While SDM using selected initializations performs better than using learned initializations. The results confirm our hypothesis on the relation of head pose and initial bounding box and demonstrate the effectiveness of our constrained initialization method.

We further choose some testing results where constrained ensemble initialization and ensemble initialization succeed single initialization and constrained ensemble initialization achieves the smallest prediction error. The visualization is shown in Figure 4.

	single initial-ization	constrainted ensemble initialization	constrained ensemble initialization with learning
Chehra	0.0522	0.0558	**0.0562**
IntraFace	0.0479	0.0521	**0.0551**
SDM*	0.0485	**0.0540**	0.0534
PO-CR	0.0606	0.0620	**0.0640**

TABLE V: AUC of Chehra, IntraFace, SDM and PO-CR using single initialization, constrainted ensemble initialization and constrained ensemble initialization through learning. SDM* means we used our implementation of SDM based on Menpo code [1].

Fig. 4: Visualization of landmark predictions using single initialization, ensemble initialization and constrained ensemble initialization. The first column shows predictions using single initialization. The second column shows results using ensemble initialization. The third column shows results using constrained ensemble initialization.

D. Discussion

Our results demonstrate the importance of selecting the right initialization locations when performing cascaded regression. This is especially important when performing landmark detection in videos with high pose variability – as is the case in 300VW.

We also demonstrate how our *constrained* initialization technique can lead to fewer "false positives" in landmark detection, leading to better overall performance.

VII. CONCLUSIONS

In our work we present a novel model initialization procedure that can help with cascaded regression based facial landmark detection and tracking. Our approach enables smart initialization of the model based based on landmark detections in the previous frame. We demonstrate the benefit of our approach on 300VW [18] – a large publicly available dataset of *in-the-wild* videos. Furthermore, we demonstrate how our approach can be used to initialize a range of modern landmark detection methods, with improved results for all of them.

REFERENCES

[1] Joan Alabort-i Medina, Epameinondas Antonakos, James Booth, Patrick Snape, and Stefanos Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 679–682. ACM, 2014.

[2] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3444–3451, 2013.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Continuous conditional neural fields for structured regression. In *Computer Vision–ECCV 2014*, pages 593–608. Springer, 2014.

[4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016.

[5] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.

[6] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollar. Robust face landmark estimation under occlusion. In *International Conference on Computer Vision*, 2013.

[7] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by Explicit Shape Regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2887–2894. Ieee, jun 2012.

[8] Grigorios G. Chrysos, Epameinondas Antonakos, Patrick Snape, Akshay Asthana, and Stefanos Zafeiriou. A Comprehensive Performance Evaluation of Deformable Face Tracking "In-the-Wild". *International Journal of Computer Vision*, 2016.

[9] Piotr Dollar, Peter Welinder, and Pietro Perona. Cascaded pose regression. In *CVPR*, pages 1078–1085, 2010.

[10] KangGeon Kim, Tadas Baltrušaitis, Amir Zadeh, Louis-Philippe Morency, and Gérard Medionni. Holistically Constrained Local Model: Going Beyond Frontal Poses for Facial Landmark Detection. In *BMVC*, pages 1–12, 2016.

[11] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[12] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, pages 679–692. Springer, 2012.

[13] David G Lowe. Distinctive image features from scale invariant keypoints. *Int'l Journal of Computer Vision*, 60:91–11020042, 2004.

[14] Georgia Rajamanoharan and Timothy F Cootes. Multi-View Constrained Local Models for Large Head Angle Facial Tracking. In *ICCV*, 2015.

[15] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 2015.

[16] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[17] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[18] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The First Facial Landmark Tracking in-the-Wild Challenge: Benchmark and Results. *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015.

[19] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.

[20] George Trigeorgis, Patrick Snape, Mihalis A. Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic Descent Method: A recurrent process applied for end-to-end face alignment. In *CVPR*, 2016.

[21] Georgios Tzimiropoulos. Project-Out Cascaded Regression with an application to Face Alignment. In *CVPR*, 2015.

[22] Georgios Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3659–3667. IEEE, 2015.

[23] Michal Uricar, Vojtech Franc, and Vaclav Hlavac. Facial Landmark Tracking by Tree-Based Deformable Part Model Based Detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 963–970, 2015.

[24] Nannan Wang, Xinbo Gao, Dacheng Tao, and Xuelong Li. Facial Feature Point Detection: A Comprehensive Survey. page 32, 2014.

[25] Yue Wu and Qiang Ji. Shape Augmented Regression Method for Face Alignment. In *ICCVW*, 2015.

[26] Xuehan Xiong and Fernando Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[27] Heng Yang, Wenxuan Mou, Yichi Zhang, Ioannis Patras, Hatice Gunes, and Peter Robinson. Face alignment assisted by head pose estimation. In Mark W. Jones Xianghua Xie and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 130.1–130.13. BMVA Press, September 2015.

[28] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *Computer Vision–ECCV 2014*, pages 1–16. Springer, 2014.

[29] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015.

[30] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015.

[31] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.