POSE INVARIANT FACIAL COMPONENT-LANDMARK DETECTION

B. Efraty, M. Papadakis, A. Profitt, S. Shah and I.A. Kakadiaris

Computational Biomedicine Lab
Departments of Computer Science, ECE and Biomedical Engineering
University of Houston, TX, USA

ABSTRACT

Facial landmark detection has proved to be a very challenging task in biometrics due to the numerous sources of variation. In this work, we present an algorithm for robust detection of facial component-landmarks. Specifically, we address the variation due to extreme pose and illumination. To achieve robust detection for extreme poses, we use a set of independent pose and landmark specific detectors. Each component-landmark detector is applied independently and the information obtained is used to make inferences about the layout of multiple components. In addition, we incorporate a multi-view representation based on an aspect graph approach. The performance of our algorithm is assessed using data from a publicly available database. The failure rate of our method is lower than that of commercially available software.

Index Terms— Facial landmarks, pose invariant, bag-of-words, component-landmarks

1. INTRODUCTION AND PREVIOUS RESEARCH

Pose and illumination invariant point-landmark identification continues to be a challenging problem in biometrics. Previous research in the field [1] suggests that if we reduce the search space for specific point-landmarks within anatomically well-defined regions-of-interest (ROI), henceforth referred to as facial component-landmarks, then existing algorithms can identify these facial point-landmarks with high-spatial accuracy. Such component-landmarks are rectangular ROIs containing an anatomical region (e.g., the mouth). Consequently, identifying point-landmarks under extreme pose and illumination variations, may become easier through the robust identification of component-landmarks under the same set of conditions. In this paper, we propose a method (which is robust to illumination and pose variations) to identify the component-landmarks. Existing methods for landmark detection that include consideration of pose variation follow two different approaches: view-based [2, 3, 4] and 3D model-based [5, 6, 7].

View-based methods use a finite partition of the continuous space of views and solve this problem independently for each view. Zhou et al. [2] used the mixture of a probabilistic PCA to solve the problem of multi-view face alignment. Su et al. [3] proposed a combined approach, where a 3D face shape model was used for view estimation to trigger view-based shape fitting. Valsar et al. [4] used Support Vector Regression to guide a search for landmarks, where the regression models were learned on multi-view facial images. Generally, view-based methods require the selection of a non-unique finite subset of orientations, which is not always uniquely defined. In our approach, this subset is defined from a combination of information from separate detectors.

On the other hand, 3D model-based methods are typically generative methods aimed at recovering both the 3D shape of the face and its “best-fit” orientation with respect to the axis of the camera that is perpendicular to the image plane. Vetter et al. [5] proposed a 3D morphable model incorporating shape and texture fit to the 2D image using optical flow. Kanade and Gu [6] proposed a three layer generative model of landmark candidates, geometric transformation and shape, that is linked through a Bayesian inference. The approach of Caunce et al. [7] involved a search within a set of view-dependent local patches that are used to update the 3D model parameters. In general, 3D-based methods require an accurate model of the face shape, which is not always available. Our approach to component-landmark identification is view-based. We seek to remove the requirement for 3D data for point-landmark localization by providing an efficient initialization to existing point-detection methods.

In this paper, pose variation is limited to the change of the yaw angle of the head ranging from $0^\circ$ (frontal view) to $90^\circ$ (left side view). The range $[-90^\circ, 0^\circ]$ can be processed in the same manner using mirror-reflected images. The extreme roll and pitch rotations are not within the scope of this paper. The contributions of our work are: (i) a method for robust component-landmark detection that works accurately for all poses varying from side to frontal view, in a variety of illuminations; (ii) the development of novel component-landmark detectors that can also be used for pose estimation, and (iii) a new component-landmark calibration procedure that is robust to inaccuracies in manual landmarking.
Channels 1-4 are the output of the IMRA filters designed to multi-scale grayscale, and three multi-scale gradient norms. Overall, we use ten channels – four IMRA, three filters \([1, 10]\), as well as multiscale gradient norm and intensity based on Isotropic Multiresolution Analysis (IMRA) wavelet filters.

For each image, a multi-channel representation is constructed. Alg. 1, respectively.

The first model is used to predict the size of landmark, and the two others are used to optimize the search for the configuration of landmark candidates. The training and detection phases of the proposed framework are outlined in Alg. 1 and Alg. 2, respectively.

2. METHODS

2.1. Overview

Our method is tuned to detect 12 component-landmarks, which contain both anatomy- and pose-related characteristics. Each landmark corresponds to a range of poses where its location variation is confined within predefined limits. For instance, the “side nose” landmark is associated with yaw range \([45°, 90°]\), whereas the “nose” landmark is associated with \([0°, 60°]\). These ranges were found experimentally based on the training error of the detector. Note that frontal and side landmarks may be present simultaneously for an intermediate range of poses. According to the presence or absence of every landmark, the space of addressed poses is divided into three views (Fig. 1).

The search region is limited to the face ROI predicted by the PittPatt face detector [8]. Each component-landmark is detected independently using a cascaded classifier utilizing Bag-of-Words (BoW) features [9]. Additionally, training data are used to learn three complementary statistical models: (i) a linear regression model which predicts the dimensions of a component-landmark as a function of the dimensions of the face ROI, (ii) Point Distribution Models (PDM) of component-landmarks encoded by their centers for each view, and (iii) the empirical distribution of each component landmark’s bounding box with respect to detected face ROI. The first model is used to predict the size of landmark, and the two others are used to optimize the search for the configuration of landmark candidates. The training and detection phases of the proposed framework are outlined in Alg. 1 and Alg. 2, respectively.

2.2. Multi-channel Image Representation

For each image, a multi-channel representation is constructed based on Isotropically Multiresolution Analysis (IMRA) wavelet filters [1, 10], as well as multiscale gradient norm and intensity maps. Overall, we use ten channels – four IMRA, three multi-scale grayscale, and three multi-scale gradient norm. Channels 1-4 are the output of the IMRA filters designed to detect singularities and edges in the direction of the intensity gradient. These singularities are spatially organized along lines which can anatomically characterize facial components. The response of these IMRA-filters does not depend on the orientation of the intensity gradient. Edge and singularity detection is accomplished by computing localized derivatives of the filtered intensity map with band- or low-pass isotropic wavelet filters adjusted for different scales. IMRA-filtering is carried out in the frequency domain by approximating the radial transfer functions to achieve a high numerical accuracy. Channels 5-7 are based on a multi-scale gradient norms using basic differential quadrature rules with appropriate spatial extent to construct features which may identify critical points and topological properties. Channels 8-10 consist of image intensity channels from various scales facilitating the extraction of texture and finer detail. All of the channels are post-processed using local intensity normalization to reduce the influence of uneven illumination.

2.3. Adaptive Feature Selection

The candidate region of any component-landmark is represented using a histogram-based feature vector. To accomplish this representation, we design a code-book following the Bag-of-Words (BoW) approach [9]. Traditionally, an unsupervised learning method is used to cluster patches from training images for the construction of a code-book; each cluster yields one code-word. However, this approach to code-book construction ignores the empirical distributions for the positive and negative samples (negative samples are all possible regions disjoint from the true location of the landmark). To construct the code-book in such a way that it follows the distribution of positive and negative sample patches, we customize the code-book so the KL-divergence of distributions for the two classes is maximized [1]. The cascade of classifiers approach [11] rejects most of the false positive candidate regions in the first few steps while keeping the classifier as simple as possible at every level. The single classifier at every level of the cascade is designed based only on the subset of the training data that could not be correctly labeled by any of the preceding levels.

2.4. Optimal Landmark Configuration Search

The cascaded-classifier is applied at every position of the sliding window until the termination level \(M\), which is the level of

---

**Algorithm 1 Training**

1. Compute multichannel representation.
2. for all Component-landmarks do
   3. Train cascaded classifier with adaptive BoW features
4. end for
5. Compute complementary statistical models.
the cascade classifier where the classification label \( L \in \{0, 1\} \) is assigned. We use posterior probabilities \( P(M = m|L = 1) \) to populate a likelihood map \( g(x, y) \) such as the one depicted in Fig. 2(a). Based on each likelihood map, we select up to three candidate ROIs per component-landmark based on their posterior probability.

Using these candidate positions (Fig. 2(b)), the landmarks optimal configuration is computed as follows: Let \( V = \{1, \ldots, h\} \) be the set of \( h \) considered views. Each view \( v \) corresponds to a particular set of visible landmarks \( \{i_1, \ldots, i_{N(v)}\} \), and each visible landmark is selected among a set of candidates. The view-specific solution \( s_v = (s_{i_1}, \ldots, s_{i_{N(v)}}) \) is encoded by a composition of decisions for each landmark. Specifically, \( s_i \in \{0, 1, 2, 3\} \), where \( s_i = 0 \) corresponds to occluded or non-detectable landmark – a “wildcard”. Every solution \( s_v \) corresponds to positions \( u(s_v) = [u_{i_1}, \ldots, u_{i_{N(v)}}] \).

Let \( \bar{u} \) represent the mean point-shape of view \( v \) learned during the training stage. The optimal view-specific solution is defined as \( s_v^{opt} = \text{argmax}_s \{Q(u(s), \bar{u})\} \), where \( Q(\cdot, \cdot) \) is the normalized Procrustes Distance between two point sets not including “wildcards”. For any given view, the solution space of this problem, is explored using a depth-first search (DFS) traversal of the directed graph with pruning based on a branch-and-bound approach. If the landmark is a “wildcard”, then its position in the final configuration is predicted using the statistical shape model [12]. The final configuration is to be chosen among all considered views (Fig. 2(c)). To accomplish this we use likelihood scores computed by the detectors for every candidate position. If \( \tilde{g}(s) \) be the mean likelihood score corresponding to solution \( s \), then \( v^* = \text{argmax}_v \tilde{g}(x^*_v) \).

### 2.5. Semi-automatic Annotation for Component-Landmarks

Unlike point-landmarks, component-landmarks are subject to high levels of ambiguity since even the human operator annotations can vary significantly. The degree of uncertainty will increase due to variations in pose. Poor accuracy in landmark annotation may influence the degree of class separability in the feature space. Thus, the precise position and size of the component-landmark has to be optimized for every image with respect to the classifier. This is accomplished by incorporating a landmark calibration method within the classifier training through the following steps: (i) train a preliminary component-landmark detector; (ii) search for the candidate region within a predefined range of positions, sizes and aspect-ratios (by applying the classifier) and select the position where the classification score is maximized; and (iii) retrain the detector based on re-annotated component-landmarks. We have observed that such an approach results in higher detection rates and tolerates inaccuracies in the initial annotation.

### 3. EXPERIMENTAL RESULTS

We use a subset of the Multi-PIE database [13] for our experiments. Two disjoint cohorts were constructed from subsets of this database using only the neutral expression. Cohort 1 (~1,000 images) was used for training, while Cohort 2 (~2,000 images) was used for testing. Examples of the poses and illuminations used in our cohorts are depicted in Fig. 3. Note that poses 05, 11 and illumination 16 (see Fig. 3) are present in the testing set, but not in the training set.

In our experiments, we assess the accuracy of the detected landmarks using the normalized overlap metric. This metric measures how much the rectangular region \( A \) of the detected component overlaps with the ground-truth region \( A_0 \):

\[
\frac{|A \cap A_0|}{\min(|A|, |A_0|)}.
\]

The results are sorted by view, since different

![Fig. 2: Optimal configuration search process: (a) likelihood maps per landmark, (b) centers of all candidate ROIs for all component-landmarks, and (c) best configuration.](image)

![Fig. 3: Subset of poses and illuminations from the Multi-PIE database.](image)
Cumulative distribution of the normalized overlap per component-landmark in: (a) nearly frontal view, (b) intermediate view, and (c) side view. Failure rates are found in the legend: our detector (numbers on the right), PittPatt detector [8] (numbers on the left).

Figure 4 depicts the cumulative distribution function of the normalized overlap on a landmark-by-landmark basis. In the legend area of this graph, we point out our detector’s failure rate based on the probability to have normalized overlap lower than 0.5. We compare these rates with the failure rates for the same landmarks detected with a commercial software [8]. This not a direct comparison, since the output of that software contains point-landmarks, which are converted to component-landmarks using the ground truth size. Note that not all the landmarks are available for comparison based on the output of [8]. We obtain comparable performance and even achieve better results for some compared landmark-view combinations.

To validate the performance of the detector on an image-by-image basis we compute the $k^{th}$ smallest normalized overlap in every image. The normalized overlap is quantized into three categories: bad (in $[0, 0.5]$), good (in $[0.5, 0.7]$), and very good (in $[0.7, 1]$). The empirical frequency of each category across view is depicted in Fig. 5. Note that side-view faces are very challenging, since they contain only few landmarks.

## 4. CONCLUSIONS

In this paper, we presented a framework for an automated component-landmark detection, that can be used for initialization of point landmark detectors, or as a stand alone module.

## 5. REFERENCES


