

REVIEW

Open Access

# A comparative study of face landmarking techniques

Oya Çeliktutan<sup>\*</sup>, Sezer Ulukaya and Bülent Sankur

## Abstract

Face landmarking, defined as the detection and localization of certain characteristic points on the face, is an important intermediary step for many subsequent face processing operations that range from biometric recognition to the understanding of mental states. Despite its conceptual simplicity, this computer vision problem has proven extremely challenging due to inherent face variability as well as the multitude of confounding factors such as pose, expression, illumination and occlusions. The purpose of this survey is to give an overview of landmarking algorithms and their progress over the last decade, categorize them and show comparative performance statistics of the state of the art. We discuss the main trends and indicate current shortcomings with the expectation that this survey will provide further impetus for the much needed high-performance, real-life face landmarking operating at video rates.

## 1 Introduction

Accurate face landmarking and facial feature detection are important operations that have an impact on subsequent tasks focused on the face, such as coding, face recognition, expression and/or gesture understanding, gaze detection, animation, face tracking etc. We define a face landmark as a prominent feature that can play a discriminative role or can serve as anchor points on a face graph. Commonly used landmarks are the eye corners, the nose tip, the nostril corners, the mouth corners, the end points of the eyebrow arcs, ear lobes, nasion<sup>a</sup>, chin etc. We prefer using the term facial component as denoting an entire facial semantic region, such as the whole region of an eye or of eyes, the region of the nose, mouth, chin, cheek, or eyebrows. Landmarks such as eye corners or nose tip are known to be little affected by facial expressions, hence they are more reliable and are in fact referred to as fiducial points. Fiducial points in imaging systems refer to marks deliberately placed in the scene to function as a point of reference or a measure. By extension, relatively stable or robust facial landmarks such as eye corners or nose tip are also called fiducial points or fiducial landmarks in the face processing literature.

Typical applications where face landmarking plays a prominent role are facial expression analysis [1,2], face animation [3,4], 3D face reconstruction [5], registration

[6,7], feature-based face recognition, verification [8-10] and face tracking [11,12], head gesture understanding [13]. Subsequent applications of landmarking could be for anonymization of facial identity in digital photos, image editing software tailored for faces, lip reading, sign language interpretation etc. Below we give more details on four these landmark dependent tasks:

- Expression understanding: Facial expressions form a visual channel for emotions and nonverbal messages, and they have a role in supporting the spoken communication [14]. The spatial configuration and temporal dynamics of landmarks provide a viable way to analyze facial expressions and to objectively describe head gestures and facial expressions. Automatic identification of action units within the framework of the facial action coding system (FACS) [15] benefits from detected landmarks and their position. Some of the approaches that use landmarks for recognizing Action Units are [1,2] and for interpreting head gestures and facial expressions are [16,17].
- Face recognition: Face recognition schemes typically locate the eye region and then extract holistic features from the windows centered on various regions of interest [18,19]. The located landmark coordinates also give rise to a number of geometric properties such as distances and angles between them [20]. In fact, anthropometrical face models, where

<sup>\*</sup>Correspondence: oya.celiktutan@boun.edu.tr  
Department of Electrical-Electronics Eng., Boğaziçi University, İstanbul, Turkey

typically the face graph nodes correspond to landmark points, combine both sources of information, the configurational and appearance sources. The graph-based methods have proved to be quite effective in many applications. One seminal work in this area is the elastic bunch graph matching technique (EBGM) [9].

- Face tracking: Most face tracking algorithms benefit from tracked landmark sequences. In the model-based group of methods [11,21], a face graph model is fitted to 60-80 facial landmarks. Face tracking is realized then by letting the model graph to evolve according to face shape parameters, facial components and geometrical relations between them. The alternative tracking approach is model-free [12,22,23] and is principally based on motion estimation. In these methods, the motion is estimated at and around the landmarks vis-à-vis some reference frame. The advantage of landmark-based tracking is that both the head motion and the facial deformations are jointly estimated. This enables us to detect and classify head gestures, head and facial emblems, interpret certain mental states as well as to extract clues for head and face animation.
- Face registration: Face registration is the single most important factor affecting face recognition performance [24]. Other applications of landmarking involve building of 3D face models from stereo, from multiple images or from video sequences where landmark points are used to establish point-to-point correspondences. For example, Jain et al. [5] and Salah et al. [6], use landmark points and the thin plate-spline (TPS) algorithm to fit a generic model to the face. This capability enables various other applications, e.g., face morphing and face animation [25]. Thus a face can be transformed into those of other individuals (inter-personal) or into different expressions of the same individual (intra-personal, e.g., a neutral face to a smiling face). In summary, face landmarking is a prerequisite for face normalization and registration whether in 2D or 3D.

The goal of this article is to present a comprehensive review of the past work on face landmarking, to categorize the multitude of algorithms, to point out novel trends, to show the performances on a comparative basis, and to understand the limitations. The article is organized as follows. In Section 2, we list the relevant facial landmarks, define the performance metrics, describe some typical feature sets and the face preprocessing steps. The major landmarking methods in the literature are categorized and reviewed in Section 3. 4 addresses a different data modality: 3D face data. Section 5 is intended as a resume of the recent trends and progress in the literature.

Section 6 describes the principal face databases used in the landmarking literature, and reports the performance results obtained with simulation results. Finally, we draw our conclusions in Section 7.

## 2 Landmarks: preprocessing, performance evaluation and challenges

### 2.1 Challenges of landmarking

Despite the plethora of articles, the quest for improved face landmarking schemes continues. On the one hand, emerging applications require that the landmarking algorithms run in real-time while operating with the computational power of an embedded system, such as intelligent cameras. On the other hand, these applications require increasingly more robust algorithms against a variety of confounding factors such as out-of-plane poses, occlusions, illumination effects and expressions. The details of these confounding factors that compromise the performance of facial landmark detection are as follows:

- Variability: Landmark appearances differ due to intrinsic factors such as face variability between individuals, but also due to extrinsic factors such as partial occlusion, illumination, expression, pose and camera resolution. Facial landmarks can sometimes be only partially observed due to occlusions of hair, hand movements or self-occlusion due to extensive head rotations. The other two major variations that compromise the success of landmark detection are illumination artifacts and facial expressions. A face landmarking algorithm that works well under and across all intrinsic variations of faces, and that delivers the target points in a time efficient manner has not yet been feasible. Figure 1 illustrates the variations that the mouth can be subjected to under different expressions and poses.
- Acquisition conditions: Much as in the case of face recognition, acquisition conditions, such as illumination, resolution, background clutter can affect the landmark localization performance. This is attested by the fact that landmark localizers trained in one database have usually inferior performance when tested on another database. A case in point is version 1 and version 2 (FRGC-1 and FRGC-2) of the Face Recognition Grand Challenge, which differ in their data collection conditions. FRGC-1 is a less challenging database collected under controlled studio conditions while FRGC-2 is an uncontrolled image sets collected under varying illumination conditions, e.g., hallways, atria, or outdoors [26] with two facial expressions (neutral and smiling). Akakin and Sankur [27] show a performance drop of about 20–30% when trained on FRGC-1 and tested on FRGC-2), and vice versa. In a more recent article,



**Figure 1** Illustration of intrinsic mouth variation over identity (upper row), expression (middle) and pose factors (lower row).

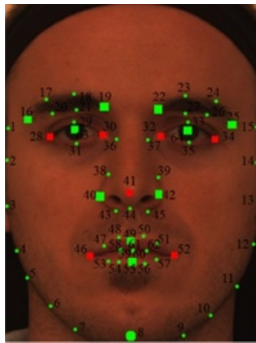
Dibeklioğlu et al. have extensively reported landmarking performances under several factors such as resolution, occlusion, expression, model choice and database [28].

- Number of landmarks and their accuracy requirements: The accuracy requirements and the number of landmark points vary based on the intended application. For example, coarser detection of only the primary landmarks, e.g., nose tip, four eye and two mouth corners, or even the bounding box enclosing these landmarks, may be adequate for face detection or face recognition tasks. On the other hand, higher level tasks, such as facial expression understanding or facial animation, require greater number of, e.g., from 20–30 to 60–80, landmarks [29,30] as well as higher spatial accuracy. As for accuracy requirement, fiducial landmarks such as on the eyes and nose need to be determined more accurately as they often guide the search for secondary landmarks with less prominent or reliable image evidence. It has been observed, however, that landmarks on the rim of the face, e.g., chin, cannot be accurately localized in either manual annotation and automatic detection. Consequently, the 17 landmark points within the face contour (4 eyebrows, 6 eyes, 3 nose, 4 mouth) are grouped together as inner landmarks and denoted as  $m17$  in the literature. We follow this tradition in our study and base most of the performance comparisons on the  $m17$  set (see Section 6). Shape guide algorithms can benefit from the richer information coming from a larger set of landmarks. For example, Milborrow and Nicolls [31] have shown that the accuracy of landmark localization increases proportionally to the number of landmarks considered and have recorded a 50% improvement as the ensemble increases from 3 to 68 landmarks.

In the final analysis, accurate and precise landmarking remains a difficult problem since, except for a few, the landmarks do not necessarily correspond to high-gradient or other salient points. Hence, low-level image processing tools remain inadequate to detect them, and recourse has to be made to higher order face shape information. This probably explains the tens of algorithms presented and the hundreds of articles published in the last two decades in the quest to develop a landmarking scheme on a par with human annotators.

## 2.2 Types of landmarks

It is convenient to consider facial landmarks in two groups, denoted as fiducial and ancillary, or primary and secondary landmarks. This somewhat artificial distinction is based on the abundance and reliability of image features aiding their detection. For example, the corners of the eyes, of the mouth, the nose tip, and sometimes the eyebrows can be detected relatively easily using low-level image features such as gradient information, cornerness or local information extracted, e.g., with scale invariant feature transform (SIFT) [32], histogram of gradients (HOG) [33], and generic information on the face morphology. These directly detected landmarks are referred to as the primary or fiducial ones, and they play a more determining role in facial identity and face tracking. The landmarks in the secondary category such as nostrils, chin, nasion, cheek contours, non-extremity points on lips or eyebrow midpoints, eyelids etc. often present scant image evidence, and the search for them is often guided by the primary landmarks. The secondary group of landmarks take more prominent roles in facial expressions, although the demarcation between these two tasks is not always clear-cut. The primary and secondary landmarks most commonly used in the literature are shown in Figure 2.



| Primary landmarks |                            | Secondary landmarks |                       |
|-------------------|----------------------------|---------------------|-----------------------|
| Number            | Definition                 | Number              | Definition            |
| 16                | Left eyebrow outer corner  | 1                   | Left temple           |
| 19                | Left eyebrow inner corner  | 8                   | Chin tip              |
| 22                | Right eyebrow inner corner | 2-7, 9-14           | Cheek contours        |
| 25                | Right eyebrow inner corner | 15                  | Right temple          |
| 28                | Left eye outer corner      | 16-19               | Left eyebrow contours |
| 30                | Left eye inner corner      | 22-25               | Right eyebrow corners |
| 32                | Right eye inner corner     | 29, 33              | Upper eyelid centers  |
| 34                | Right eye outer corner     | 31, 35              | Lower eyelid centers  |
| 41                | Nose tip                   | 36, 37              | Nose saddles          |
| 46                | Left mouth corner          | 40, 42              | Nose peaks (Nostrils) |
| 52                | Right mouth corner         | 38-40, 42-45        | Nose contours         |
| 63,64             | Eye centers                | 47-51,53-62         | Mouth contours        |

**Figure 2** *m7* landmark set includes squares representing the primary (first order) landmarks. *m7* landmark set consists of the most fiducial points represented by red squares. Green dots the secondary (second order) landmarks, totally 64 landmark points.

### 2.3 Landmarking performance

One can define two different metrics to evaluate landmarking performance: (i) ground-truth based localization error; (ii) task-oriented performance. For ground-truth based localization error, a straightforward way to assess landmarking performances is to use manually annotated ground-truths. For task-oriented performance, one can measure the impact of the landmarking accuracy on the performance scores of a task.

A straightforward way to assess landmark detection and landmark localization performances is to use manually annotated ground-truths. If the ground-truth positions are available, the localization performance can be expressed in terms of the normalized root mean square error (NRMSE). NRMSE can be computed per landmark or NMSE figures can be averaged over all the landmarks to produce a global precision figure. The normalization is typically done with respect to IOD: Inter-Ocular Distance, which is defined as the distance between the two eye centers. Normalizing landmark localization errors by dividing with IOD makes the performance measure independent of the actual face size or the camera zoom factor.

One can declare a landmark to be detected whenever the localization error remains below a suitably chosen error threshold,  $Th$ . The landmark errors are assumed isotropic, so that one can conceive around each ground-truth landmark a detection circle with radius equal to the error threshold. If the Euclidean distance of the estimated landmark is below the threshold, the landmark is considered as detected; otherwise, whatever the value of the localization error, it is declared as a missed landmark. A detection circle is illustrated in Figure 3. A nice way to illustrate the detection performance is to plot the percentage of times a particular landmark is detected within a given error radius. In Figure 3, the abscissa denotes the error radius while the ordinate is the empirical probability of landmark detection. The allowed error radius (detection threshold)

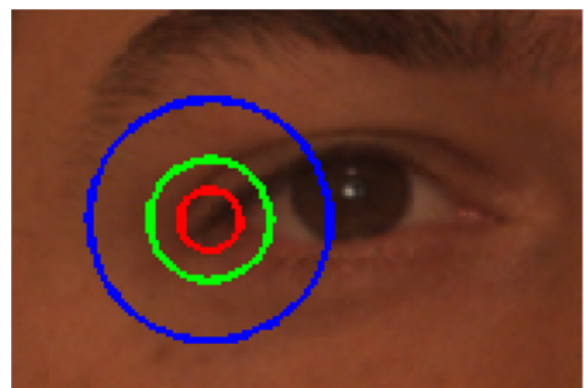
is taken as some percentage of the inter-ocular distance IOD, typically 10% or below of IOD.

The localization precision is thus computed as the Euclidean distance  $d(., .)$  between the ground-truth coordinates,  $(x, y)$ , and the estimated coordinates,  $(\tilde{x}, \tilde{y})$ , normalized by  $d_{norm}$ , the IOD. The error is given as

$$\delta_i^k = \frac{d\{(x_i^k, y_i^k), (\tilde{x}_i^k, \tilde{y}_i^k)\}}{IOD}, \quad (1)$$

where the superscript  $k$  indicates one of the landmarks (e.g., eye corner, nose tip) and the subscript  $i$  is the image index.

Landmark detection statistics can be characterized by the exceedance probability of the localization error. A general agreement in the literature is that  $\delta_i^k < 0.1$  is an acceptable error criterion so that a landmark is considered detected whenever it is found within proximity of one tenth of the inter-ocular distance from its true



**Figure 3** Detection probability of the left eye outer corner versus normalized error. Concentric circles denote error ranges with radii 0.05 (red), 0.1 (green), and 0.2 (blue) times IOD, respectively.

position. More specifically, we calculate the per-landmark performance:

$$P(k) = 100 \frac{\sum_{i=1}^I [i : \delta_i^k < Th]}{I} \quad (2)$$

where  $[i : \delta_i^k < Th]$  is the indicator function assuming 1 if the deviation is smaller than a threshold, otherwise its value is 0, and  $I$  denotes the number of test images. The overall performance is averaged over all landmark types

$$P = 100 \frac{\sum_{k=1}^K \sum_{i=1}^I [i : \delta_i^k < Th]}{K \times I}. \quad (3)$$

A goal-oriented landmarking measure could be its impact on the performance of tasks. Some instances of goal-oriented applications based on landmarking are face registration algorithm, expression classification as in [28] or fitting of the active appearance model (AAM) algorithm as in [34], and gesture recognition as in [16,35]. The landmarking accuracy on the performance of the registration, expression classification, gesture recognition and AAM fitting algorithms, respectively, would be goal-oriented measures of landmarking.

## 2.4 Preprocessing for landmark extraction

There is always some preprocessing before a method engages in landmark detection. Typical of these steps are the following: illumination artifact removal, modest geometric corrections, segmentation of the face, use of color information.

### 2.4.1 Illumination compensation

The detected face region is subjected to illumination compensation, which can be operated pixelwise, locally or globally. One example of pixelwise normalization is CSDN: center-surround divisive normalization [36], where each pixel is divided by the mean value of a block around it; another example is rank filtering where pixels in the surrounding block are ranked, and the central pixel is simply assigned its rank value and all such assignments finally stretched to the  $[0, 255]$  interval. Local normalization can be attained via filtering with Laplacian of Gaussians or using a facet model as in [37]. Finally, the prototypical example of global normalization is histogram equalization.

*Use of geometry:* The task of landmark localization is aided by the knowledge of the geometric relationship (distances, angles etc.) between landmarks and overall shape characteristics. This knowledge can be converted to a set of rules and/or can be expressed as a set of statistics of point-to-point distances and angles subtended by local ensembles, e.g., triples of landmarks. The eyes and sometimes the mouth can be found via an algorithm like Viola-Jones [38], Gabor filters [39], projection histograms [40], specifically trained SVMs [41],

or else. Once a few facial components are detected, e.g., the eyes and mouth, geometry information can be used to initialize the search for the remaining ones in a reduced search area. Geometric constraints also help the post-processing stage where landmarks are geometrically verified. For example Shih and Chuang [39] initialized the mouth at one IOD below the eyes, nostrils within  $0.6 \times \text{IOD}$  below and eyebrows within  $0.4 \times \text{IOD}$  above etc. If certain landmarks are missing or if the detected landmarks or face components do not to satisfy given reliability criteria, they can be recovered or their search re-initialized via the geometric face model.

### 2.4.2 Face segmentation

A commonly occurring theme is the heuristic segmentation or compartmentalization of the face encompassing target regions of interest [39,41-43]. For example, the face is partitioned with a grid structure resulting in two or more horizontal and vertical stripes. This helps to delineate the search areas so that, e.g., the eyes are searched in the northeastern and northwestern corners while the mouth is searched in the southern sector [44]. A popular method to segment the face is to use projection histograms [40,45]. The relatively darker eye and mouth regions cause dips in the histograms and the corresponding bands are used to initialize the search for the eyes and mouth. Pitas and Tsekeridou [46] take advantage of the mirror symmetry of the face and use vertical and horizontal histograms to initialize the location of face components. Most of the recent study, e.g., [47] however, use training images to learn the a priori location of the target landmark within the bounding box of the detected face.

### 2.4.3 Role of color

Color information, mostly in the pre-Viola-Jones era, has been used for face and mouth detection. For example, in [48], Hsu et al. proposed a face segmentation method based on skin labeling in the non-linear YCbCr color model and the connected components analysis. Within the face region, any non-skin colored blob is regarded as a candidate for eyes or mouth. In a companion study, in [49], color information is used in the energy functional of the snakes [50] or to assist in the initialization step and fitting process of ASM [51]. There have been also a number of studies, e.g., [52], to detect lips automatically based on their color properties.

## 3 Review of face landmarking methods

The plethora of face landmarking methods in the literature can be categorized in various ways, for example, based on the criteria of the type or modality of the observed data (still image, video sequence or 3D data), on the information source underlying the methodology



(intensity, texture, edge map, geometrical shape, configuration of landmarks), and on the prior information (e.g., anthropometrical data), if any used.

The goal of any categorization attempt must be to lead to a better understanding of the commonality and differences of the approaches, and to extract a sense of where the state-of-the-art is heading. Despite the difficulty of finding clear-cut distinctions since algorithms often share techniques common to more than one category, nevertheless we have found it useful to categorize them based on the type of information used and on the specific methodology [53,54]. In a previous such attempt, Phimoltares et al. [53] have used the five categories of geometry-based, color-based, appearance-based, edge-based and motion-based landmarking algorithms.

We believe that there are two basic categories of facial landmark detection methods: model-based methods and texture-based methods. Model-based methods, also known as shape-based methods, consider the face image and the ensemble of facial landmarks as a whole shape. They learn “face shapes” from labeled training images, and then at the test stage, they try to fit the proper shape to an unknown face. The second category, texture-based methods, also known as non model-based methods, aim to find each facial landmark or local groups of landmarks independently, without the guidance of a model. In these methods the shape information may still be invoked, but at a later stage for verification.

These two broad categories of landmarking methods can each be further split into two sub-categories. The model-based methods can be split as explicit methods, of which prime examples are ASM and AAM, and as implicit methods, for example, algorithms using a neural network applied to the whole face. Similarly, the texture-based methods can be discussed under the sub-categories of transform-based methods, e.g., Gabor filters or HOG features, and template-based methods. Figure 4 illustrates this categorization. Note that the transform methods can further be split into linear transform methods, like principal component analysis (PCA), independent component

analysis (ICA), Gabor transform, and nonlinear transform methods like Kernel PCA (KPCA), local linear embedding (LLE) etc. However, in this study subcategorization at this detail was not warranted. At this stage, to preclude any misinterpretation, we have to emphasize that the shape and texture approaches are not mutually exclusive. As a case in point, the shape-based methods do also utilize the local texture information to guide the model shape to fit; conversely, most of the texture-based methods eventually use some shape information, e.g., at a later stage for verification. The shape or texture categorization in this article puts into evidence the predominant source of information in marking. Thirdly, it would be possible to extend the taxonomic tree by including the category of 3D landmarking methods (Section 4). However, we consider 3D as a different data modality rather than a methodological category. In fact, other interesting data modalities that we could consider would be infrared images, photo-sketch based images etc.

### 3.1 Texture-based methods

The texture based category of methods will be considered in two classes, namely, transform-based and template-based. In the transform-based schemes, a window scanning the image has its content transformed into a feature vector and this feature vector is compared with the learned patterns. In the template approach, a landmark template or a set of landmark templates scan the image to identify the target landmark according to the strength of the template matching response. An overview of the texture-based methods is given in Table 1.

#### 3.1.1 Transform-based methods

Pioneering examples in this track are the modular PCA method and eigenfeatures, which are essentially the eigenface approach specialized to facial components. Pentland et al. [18] derive eigenmouths, eigeneyes and eigennoses, which coupled with eigenfaces result in good recognition in multi-pose face databases. The authors also point out

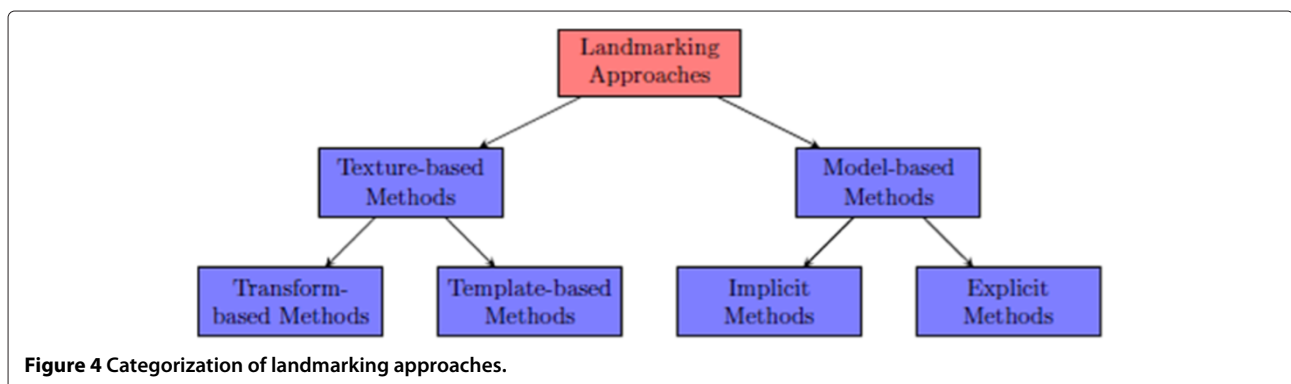



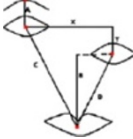






Figure 4 Categorization of landmarking approaches.

**Table 1 An overview of the texture-based face landmarking algorithms**

| Work                            | Highlights of the method   | Domain knowledge used  | Landmark types  |
|---------------------------------|--|--|---|
| Yuille et al. [72], 1989        | Using image saliencies of the face components, geometrical templates are developed consisting of arcs and circles. Eye template consists of a circle for iris, two parabola sections for eye contours, two center points for the white sclera.   | Descriptive information of the eye and mouth geometries.   | Eye, iris and mouth contours.<br>              |
| Pentland et al. [18], 1994      | Extension of the eigenface approach to eigenmouth, eigeneye and eigennose. Multiple eigenspaces mitigate variations due to pose. Face-ness, mouth-ness etc. are assessed based on the concept of distance from corresponding (eye, mouth, nose etc.) eigenspace.                       | None.  | Mouth, nose and individual eye components.<br> |
| Vukadinovic & Pantic [44], 2005 | GentleBoost templates built from both gray level intensities and Gabor wavelet features. A sliding search is run with templates over twenty face regions.  | Face initially divided into search regions on the basis of IOD vis-à-vis the detected eyes. In addition horizontal and vertical projection histograms and symmetry of the frontal face are used. | 20 landmarks.<br>                              |
| Arca et al. [41], 2006          | Face is detected with skin features, and eyes are located using SVM. Facial components are extracted using parametric curves specific to each component as in [72], and facial landmarks are traced on these curves.   | Various facial component heuristics such as the vertically alignment of the eyes, the mouth is centered with respect to the eye positions etc.   | 16 landmarks<br>                              |
| Zhang & Ruan [73], 2006         | Rectangular eyes, mouth and nose templates resulting from averaging several instances used for detection. Geometrical templates consisting of arcs and circles are fitted to components for detailed modeling.   | Eye and mouth geometry.  | Eye, iris and mouth contours.<br>            |
| Akakin & Sankur [16,27], 2007   | Templates based on 50% of block DCT features (block size $0.4 \times \text{IOD}$ ) scan the image and SVM score map is obtained. Initial combinatorial search decides for 7 fiducial landmark, and the rest of the landmarks are predicted and locally tested with their DCT features. | Landmark distances and angles are learned, modeled as Gaussians and the information embedded in a graph.   | 17 landmarks.<br>                            |
| Ding & Martinez [68], 2010      | Face components are found via Subclass Determinant Analysis, where multiple models for the target component, eyes and mouth are developed; the context is the subspace representation of the regions surrounding the components.   | Estimated positions of the face components within detected face boxes.   | Eyes and mouth components.<br>               |
| Valstar et al. [70], 2010       | SVRs are trained to predict the landmark locations using RoI samples. The search is regularized via a Markov network to exploit the learned spatial relationships between landmarks.   | A priori probability map of the likely locations of seven fiducial landmarks and the locations of 15 less fiducial landmarks vis-à-vis the first seven.  | 20 landmarks as in [44].<br>                 |

to the higher tolerance of the eigenspace approach to geometric variations as compared to simple template matching, and claim a method for view-independent facial component detection. Other instances of the

appearance-based category extract features using Multi-resolution Wavelet Decomposition [55-57], Gabor wavelet transform (GWT) [44,57-60], discrete cosine transform (DCT) [27,61], and independent component

analysis (ICA) [62]. For any given landmark, after its features are extracted, a classifier is trained to convert the feature vector into a likelihood score.

#### *Gabor transform*

Gabor wavelet transform can produce effective features as these wavelets are known to generate biologically-motivated convolution kernels [9]. To this effect, a bank of filters is created by using Gabor kernels in different orientations and frequencies (scales), and then convolved with the data to form a Gabor space. In this vein, Smeraldi and Bigun [59] developed a bio-inspired scheme where Gabor features were computed on a retinotopic sampling grid on the face. For uniform frequency coverage, they used modified Gabor filters, where shifts of Gaussians on the log-polar frequency plane correspond to rotations and scaling. The facial components are found by displacing the retinotopic grid and evaluating the output with SVM classifiers. Similarly, Vukadinovic and Pantic stacked Gabor responses at eight orientations and six scales into a feature vector [44]. This feature is then used to train a Gentle-boost classifier for each landmark type within its own region of interest (RoI).

Ersi and Zelek [60] conjectured that facial components must have higher entropy as compared to the rest of the face, and they initialized the search regions with a high entropy threshold. Subsequently, facial components are verified by using the combined features of Gabor coefficients and the local entropy (entropy of the search window). Two- or multi-tiered approaches are common search strategies [57,63]. A two-level hierarchical Gabor wavelet networks (GWNs) is presented in Ferris et al. [57]. Here a GWN denotes a constellation of 2D Gabor wavelets that are specifically chosen to reflect the object properties and together with its location they constitute a node in a tree representation. The hierarchy consists of a set of GWNs that are organized in child node-parent node relationship, which may differ in position, scale and orientation. The first-level network, trained for the whole face, yields orientation and scale information of the face as well as approximate landmark locations. The set of second-level networks, trained for each landmark separately, yield refined landmark locations. Duffner and Garcia [63] also used a neural architecture in such a hierarchical way where the search area is restricted by the preceding step.

#### *Discrete cosine transform*

Salah et al. [58] presented another coarse-to-fine approach where they first search on a lower resolution image for coarse landmark locations, and then refine them on the original resolution image. In a comparative analysis, they observed that DCT features perform slightly better

than Gabor features, both using SVM classifiers [58,64]. DCT coefficients have also proved to work surprisingly well as low-level features leading to high localization performance, and furthermore they offer the advantage of the existing DCT implementation. In [61], Zobel et al. used DCT features in a probabilistic structure to detect facial components where the spatial dependencies between landmark points are statistically modeled. More specifically, the length of rays emanating from the face center and pointing to the eyes and mouth as well their angles subtended were modeled as Gaussians. Akakin et al. [27,58] generalized this idea and used a probabilistic graph-based framework as a post-processing tool to correct erroneous estimates and recuperate missing landmarks.

#### *Independent component analysis*

In [62], Antonini et al. resorted to ICA features to exploit higher order dependencies in images. They initialize candidate locations with Harris corner detector and proceed to extract ICA features within  $32 \times 32$  windows at these corners. The resulting feature vectors are classified with SVM to result in 10 landmarks. It is interesting to note that the ICA method applied to Gabor features, resulting in the so-called independent Gabor features (IGF), improves the performance over the Gabor only method [64].

#### *Landmark initialization heuristics*

Since the costly part of appearance-based methods is the brute-force search for the landmarks, an efficient method to restrict the search area is the use of vertical and horizontal gray-value projections. Projections are simple to implement, while being at the same time quite effective in determining a first coarse estimate of feature positions. Brunelli and Poggio [65] have performed edge projection analysis by partitioning the edge map in terms of horizontal and vertical edge directions. Other researchers found out that more exact results can be obtained by applying the projection analysis on the intensity image, since most faces have fairly smooth contours faces [39,40,44,66,67]. These approaches use histogram dips caused by the gray level variations of the face components, which for the eyes and mouth regions tend to be darker than the skin level.

Similarly, Ding and Martinez [68] define two patches enclosing the eyes through a statistical analysis of eye locations in the training images. They resort to subclass discriminant analysis (SDA) [69] to combat the variability of the appearances of the eye components. Thus, the gray-level distributions of the patches centered on the components and of those in their vicinity of the target regions are modeled with K-means clustering, where the



optimal value of  $K$  is separately determined for the target and surrounding non-target patches. While eye location information aids in the detection of the remaining landmarks, chin proves especially difficult, and consequently it is deduced with the aid of a quadratic curve fitted to the available face boundaries [68]. In a more recent study, Valstar et al. [70], after finding the face box, model the prior probability of the  $x$ - and  $y$ -position of each facial point relative to the coordinate system of the detected face. Thus the  $x$ - and  $y$ -coordinates of the landmarks are described by a set of bivariate Gaussians, with means and variances learned using training images.

### 3.1.2 Template-based methods

The main differences between template-based and transform-based category are the matching procedure and the representation of facial features. Transform-based methods reduce the observed image patch to a feature vector, which is then classified; template-based methods calculate the matching scores.

#### *Fixed templates*

In fixed template methods, a face object (component or landmark) is detected if it responds strongly when convolved with the template mask, and the highest scoring point on the face is declared as the target location. The template can be obtained using adaptive matched filter techniques [65]. However, the weakness of the template-based methods is their scale and pose dependency, since cross-correlation operation is highly sensitive to geometrical distortions. To mitigate scale dependency of these methods, Poggio et al. proposed a scheme where target locations are investigated at multiple resolutions [19]. In a follow-up study focused on face recognition robust to poses, Heisele et al. [71] used 14 component detectors based on component SVMs. The geometrical configuration of these components were analyzed on a second level in order to complete the face detection and recognition tasks.

#### *Deformable templates*

Deformable templates are proposed to cope with the limitations of fixed template matching. A pioneering study in this direction was presented by Yuille et al. [72]. Accordingly, a deformable parametric template is made evolve by internal and external forces. The template is attracted to salient features such as peaks, valleys and edges in the intensity images. For example, eye and lip templates are first extracted via morphological filters and the energy function resulting from internal and external forces is minimized with a gradient descent algorithm [72]. In a recent study [73], Zhang and Ruan combined

the fixed and deformable templates, such that first, fixed templates are used to locate a rectangular ROI around the face components, and then deformable templates are used to extract the contour of the component. Notice that template or transform techniques are often used in model-based algorithms as well as part of low-level image processing.

In summary, texture-based methods generate landmark candidates independently from each local detector. These result in a score surface for each landmark on the test face, whether the score is the outcome of the matched filter or of the classifier, e.g., SVM. The peaks on the score surface, judiciously chosen in the light of a prior model for face geometry form the landmarks. Algorithms often attempt to solve the combinatorial search problem using various heuristics or invoking learned face models. The enforcement of the prior information plays the role of a regularizer.








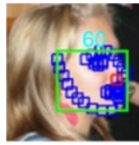
### 3.2 Model-based methods

Shape-guided or model-based methods consider the whole face and the ensemble of landmark as an instantiation of a shape. Of the two sub-categories of model based methods, the explicit model-based methods are by far more popular, while there are only a few research articles in the alternative implicit methods. Nevertheless we will discuss it briefly for the sake of completeness. Major model-based methods are listed in Table 2.

#### 3.2.1 Implicit model-based methods

Implicit models based methods use models without state information: unstated models. Methods that use pixel gray levels as input of a neural network to detect multiple facial landmark try to learn implicitly the spatial relations between landmark points. Search with genetic algorithms can also be considered under this category. For example, Cesar et al. [74] use inexact graph matching to discover landmarks. The model image is segmented at and around landmarks while the test image is oversegmented, e.g., using watershed algorithm, hence it contains a much larger number of segments as compared to the training (model) images. Landmarking of the test image, in other words, labeling of the segments as belonging to a landmark location is carried using inexact graph matching. The global dissimilarity function between the two graphs is minimized using randomized tree search and genetic algorithms. Ryu and Oh [75] segment the face, and using a number of heuristics about the geometry of facial components of interest, develop a face template based on genetic algorithm, and apply multilayer perceptrons as nonlinear templates at the landmark level.

**Table 2 An overview of the model-based face landmarking algorithms**

| Work                             | Highlights of the method  | Domain knowledge used   | Landmark types  |
|----------------------------------|---|---|---|
| Leung et al. [76], 1995          | Face image is Gaussian filtered at multiple orientations and scales. This process provides a set of candidate landmarks. Each possible configuration of candidates is validated through random graph matching.  | The geometrical relationship between landmarks is expressed with a probabilistic model, which reduces the matching complexity and eliminates irrelevant points. | Eye centers and nose.<br>  |
| Wiskot et al. [9], 1997          | A labeled graph is constructed where links are the average distances between landmarks and where nodes represent 40-dimensional Gabor jets at candidate locations. The face graph is elastically deformed toward the query face.  | Multiple face graphs capture head rotations and bunch graphs capture the various appearances.   | An example graph:<br>  |
| Cootes et al. [79], 1998         | AAM, a generalization of ASM, jointly models the shape and texture variation of the fiducial points. The main goal is to find the appropriate model parameters that minimize the difference between the query and the model face.   | PCA models of both texture and shape.   | An example of fitting:<br>   |
| Cristinacce et al. [80,81], 2003 | Multiple landmark detectors are run on the face and locate the initial landmarks. Then, two steps are repeated until convergence: First, estimated locations are improved by boosted regression; second, shape model is fitted to the updated landmark locations.                   | Configurational constraints are applied to eliminate false positives as well as to recover missing landmarks.   | 17 landmarks: eye, eyebrow, nose, mouth and chin.<br>             |
| Cristinacce et al. [83], 2008    | Local templates per each landmark type are combined into a geometrical configuration. The estimated locations are updated by a shape-driven search.   | Learned global shape model to avoid non-plausible face shapes.  | 22 landmarks.<br>  |
| Milborrow and Nicolls [31], 2008 | Enhancements on ASM such as stacking of two ASMs for better initialization, 2D profile search for individual landmarks etc.   | Learned profile models for the individual landmarks and learned global shape model via PCA  | 76 landmarks.<br>  |
| Belhumeur et al. [106], 2011     | A local detector collects SIFT features and landmark-specific SVMs output landmark likelihoods. A Bayesian framework unifies the local evidences into a global shape.   | Anatomical and geometrical constraints on facial landmarks derived implicitly from the exemplars.   | 29 features.<br>   |
| Zhu & Ramanan [99], 2012         | Local and global information merged from beginning via tree-connected patches covering the landmarkable zones of the face. Patches represent HOG features while global shape is imposed via quadratic springs between them. The maximum likelihood setting of the tree is searched. | Linearly-parameterized, tree-structured pictorial structure of the landmark rich parts of the face.   | 68 landmarks for frontal and 39 landmarks for profile faces.<br> |

### 3.2.2 *Explicit model-based methods*

Most of the explicit methods can be subsumed under the topics of graph methods and active appearance methods.

#### *Graph methods*

The study of Leung et al. [76] based on random graph matching is one of the first methods in this category. They start with a set of candidate points by convolving the image with multi-oriented and scaled Gaussian derivative filters. In order to validate the spatial arrangement of the landmarks, each configuration is tested by matching it against a graph obtained from training data. In these graphs, the normalized mutual distances between landmarks are assumed to be Gaussian distributed. This probabilistic model helps also to eliminate irrelevant configurations and reduce the matching complexity.

A seminal study in graph fitting was presented by Wiskott et al. [9], called elastic bunch graph matching (EBGM). In this model, nodes of the graph are characterized by Gabor jets, that is, Gabor wavelet responses at several scales and orientations, all stacked as a vector. The Gabor jet energies help to register the nodes of the graph to the face landmarks where the graph similarity measure takes into account both the magnitude and phase information of the jets. This method deforms elastically the face graph depending upon collective Gabor responses while preserving its geometric configuration. In the same vein, Zhu and Ji initialize a set of 28 landmark points using a face model [77], scaled to the detected face size. In the first iteration, landmarks are anchored at the eyes, which are detected via the Viola-Jones algorithm [38]. The initial mesh estimation is successively refined using Gabor jets and a slightly modified version of EBGM (using grid search based procedure for matching the phase coefficients). Any abnormal deviation of facial landmarks is corrected via PCA subspace model. One disadvantage of these configurational models such as EBGM or AAM is that they need a good initialization, which is not straightforward for unknown head poses. To overcome this limitation, Wiskott et al. [9] trained graphs corresponding to profile, half-profile and frontal faces.

#### *Two-tier graph methods*

In the studies of [68,78], statistical decision theory is compounded with geometry information for additional robustness. Most likely-landmark locator (MLLL), proposed by Beumer et al. [78], can also be thought as a variety of Viola-Jones algorithm. MLLL aims to maximize the likelihood ratio of a set of points to be in the proximity of a landmark versus the negative case. This initial localization step is followed by a shape correction method

based on PCA subspace projection and elimination of false positives.

In this vein, Akakin and Sankur [27] continue the two-tier landmark-refinement tradition and employ a probabilistic graphical model (PGM). The initial landmark estimates are found via landmark specific SVMs operating on DCT masks. Each landmark neighborhood is described by selected zonal DCT coefficients. The arcs between the nodes of the PGM (landmarks) and the subtended angles are modeled as Gaussian spring forces with parameters learnt during a training phase. Obviously, the spring forces toward nearer landmarks are stronger since one expects that the corresponding anthropometric variability will be smaller and those linking the more distant landmarks will be weaker. For example, the left outer eye corner would be tightly coupled to left eye inner corner, but more loosely coupled to right eye corners, to mouth corners or to the nose tip. The PGM accepts  $m$  candidates for each of the  $k$  landmark points ( $m$  is usually a small number, a function of the specific landmark), resulting in a combinatorial search. The  $n$  points, composing the best configuration, are called the support set and they do not necessarily cover all the landmarks. This support set is used as anchor points for adapting the graph to the actual face. Any landmark missing due to occlusion or poor data, ( $k - n$ ) points, is estimated using back-projection of the graph [64]. In [16], the authors increase the number of detected landmarks to 17 and used them for the purpose of facial expression and head gestures analysis.

#### *Active shape and appearance models*

The most important paradigm in the model-based category consist of the active shape model (ASM) and AAM varieties and their various descendants. In ASM, the deformable objects (i.e., faces) are represented by a set of fiducial points, which are found with feature detection methods. Configurational shape variations are regularized by PCA so that the face shape can only deform in controlled ways learned during training. In the same vein, AAM is proposed to impose jointly the constraints of shape variation and texture variation [79]. In AAM, the shape and texture are combined in the PCA subspace such that PCA coefficients are jointly tuned to account for the geometry and texture differences from the mean face. Recall that, in contrast to ASM, AAM is a generative approach in that by adjusting the model parameters, the shape and texture of plausible new faces can be generated. In a sense, a model face is morphed into a target image so as to minimize the model fitting residual.

Cristinacce and Cootes have also proposed a Shape Optimized Search algorithm where the feature responses corresponding to the landmark shape models are learned using the ASM [80]. Three types of landmark features are

used, namely, (i) Gray-level information ( $15 \times 15$  windows reduced with PCA), (ii) Orientation maps resulting from Sobel edge operator, (iii) Features resulting from a boosted classifiers, e.g., [38]. The concatenation of these three feature sets forms the landmark shape vector. The shape vector is then used in a shape-guided search to select the best candidate among possible configurations while the positions of missing points are predicted by the shape model. The Shape Optimized Search method, where in effect templates are learned under configurational constraints, outperforms AAM approach [79]. Several variants of ASM/AAM have recently been proposed. One of them is the boosted regression ASM [81]. The main difference of this variant is that it uses landmark detectors based on Haar features and boosted classifiers [38] instead of eigen models. In another variant [82], separate boosted cascade detectors, one for each landmark, model shapes implicitly by learning the pairwise distribution of all true feature locations. This approach can be thought as a combination of multiple detectors, called pairwise reinforcement of feature responses (PRFR). Finally an AAM which models edge and corner features instead of normalized pixel values is used for refinement.

An important step forward from the AAM algorithm is the CLM: constrained linear model algorithm [83]. Despite inheriting some of the important tools of AAM, CLM differs from AAM because it is not a generative model for the whole face, instead it produces iteratively landmark templates and applies a shape-constrained search technique. Like AAM, CLM also profits from labeled training set. The idea of template update was used by the authors in their previous study [79]. In [83], this idea was developed further so that the position vectors of the landmark templates are estimated using the Bayesian formulation. The posterior distribution in the Bayesian formula incorporates both the image information via template matching scores and the statistical shape information. Thus, new landmark positions are predicted in the light of the image and the joint shape model, and then templates are updated by sampling from the training images. The optimization search is instrumented via Nelder-Mead simplex algorithm.

Another important contribution to the ASM methodology was made by Milborrow and Nicolls [31], and their software, called standard ASM (STASM) is practically one of the standards, widely adopted by the community. Their starting point is the original work of Cootes and Taylor [84], and they modify several of the steps such that cumulatively the resulting algorithm—STASM—works roughly 60% better, in that the average landmarking errors for  $m17$  decreases from  $0.08 \times \text{IOD}$  to  $0.05 \times \text{IOD}$ . The three improvements consist in using two dimensional profiles

while searching for landmark updates, second, in adapting the shape model along the progress of the iterative search by varying the number of shape eigenvectors, and finally in running two search steps in cascade to recover from fatal starts. A recent addition to the constrained local models is the study of Saragih et al. [85] where the authors bring a probabilistic interpretation to the optimization of the statistical shape model parameters.

#### 4 Landmarking of 3D faces

Although most of the methodological advances in face landmarking has been realized on 2D images, the interest in processing 3D face images is rapidly increasing due to the wider availability of 3D cameras, e.g., Kinect sensor device, the evolution of 3D television and video. A recent review article on 3D human face description [86] traces the history of the use of landmarks from anatomical studies to aesthetic concerns, from face recognition to anthropometric measures for face correction.

The anatomical landmarks used in 3D are the same as those used in 2D images; while 2D image landmarking uses the gray-level features, 3D benefits from surface curvature features. There are, of course schemes that benefit simultaneously from 2D texture and 3D curvature information since 3D imaging devices provide also registered 2D optical images. One advantage of landmarking in 3D is that it enables alternate processing techniques for landmarks since there are multiple ways of representing 3D face data. For example, point clouds, depth maps, multiple profiles, voxels, curvature and shape index [87] have been used for face recognition, and these have not yet fully exploited for landmarking. A more important advantage is that it can potentially mitigate some of the limitations encountered in 2D landmarking. Recall that 2D landmarking becomes very sensitive to pose variations beyond  $20^\circ$  tilt and/or yaw, and it suffers also from illumination effects. In this sense, 3D face data has the promise of filling the performance gap in the presence of severe lighting and pose variations. The downside is that 3D face raw data demands substantially more preprocessing as compared to 2D. For example, the face surface must be smoothed, spikes and discontinuities removed, and gaps filled in.

##### 4.1 The use of heuristics

3D face data provides the heuristics of nose tip defined as the closest point to the range camera. The nose tip is found to be a reliable and easily detected landmark, barring poses with excessive tilt and/or yaw. In fact, many studies in the literature have exploited this simple heuristic. Lu and Jain estimated jointly the yaw angle of the face and the nose tip by exhaustively searching over quantized sectors [88]. The remaining fiducial landmarks (inner eye corners, mouth corners and chin)

are estimated using the face orientation information, corneriness feature from intensity image and shape index from the range image [89]. Note that their method [88] allows a wide pose angle range extending from  $-90^\circ$  to  $90^\circ$ . Dibeklioglu et al. [90] introduced a heuristic approach to increase the performance of 3D landmarking. They extracted the relevant regions by thresholding the difference between the Gaussian curvature and the mean curvature images. This difference image highlights the facial landmarks and its higher peak locations correspond to candidate regions. They determined a circular interest region that embraces all the components left after thresholding the difference image. They also profited from these heuristic rules on the curvature difference image to determine a circular RoI that embraces all the components and the nose tip robustly, both even under occlusions and severe pose variations.

#### 4.2 Average face model

Gökberk et al. [87] solved the problem by initializing the landmarks of the test face with the landmarks of an average face model (AFM), the two being aligned initially via iterative closest point (ICP) algorithm. The landmark positions are then searched and refined using such shape descriptors as Gaussian curvature, mean curvature, surface normals and relative distances to the facial symmetry plane.

#### 4.3 Surface curvature

Akagündüz et al. [91], inspired by the scale-invariant feature transform (SIFT), described the facial surface with the mean and Gaussian curvatures. The curvature data is calculated at many scales using a Gaussian pyramid and then binarized via thresholding. This yields a 3D curvature field, with two spatial dimensions (UV) and one scale dimension (S). The facial components (chin, nose and eye pits) are then identified using connected components in the UVS space, and their geometrical relationships subsumed by a graphical model. Segundo et al. [92] also use the curvature field the landmarks that are the least affected by expressions, namely the nose tip, eye corners and nose corners. Using biquadratic approximations to the local surface, the Gaussian and mean curvature are computed to identify the peaks and pits. For example, eye corners present a pit-like surface, and the nose tip presents a peak-like surface. The coordinates of the landmarks are found by a number of heuristics, such as the projections of the depth information reliefs.

Nair and Cavallaro [93] use a different approach, that of point distribution model (PDM for face detection), registration, landmarking and description. The PDM actually represents the face shape including the required landmarks, as well as the statistical information of the shape variations. Once a statistical model of PDM is obtained

using a training set (49 ground-truthed landmarks are used), one proceeds to fit this model to the actual faces. The fitting process is guided by the curvature-based feature map characteristic of the faces. The model is initialized with fiducial landmarks such as eye corner (endocanthus, exocanthus) curvatures. These vertices generate the remaining candidates, and they eventually settle on the landmark as a result of model transformation with the minimum deviation from the mean shape, while respecting the constraints of subspace shape.

Conde and Cabello [94] used spin images to characterize the local surface around the landmarks. Spin images can be thought of as a 3D-to-2D mapping such that each patch is characterized by relative distances of the surface points to a reference point, some sort of distance histogram. The search space is reduced by selecting the areas with higher mean curvature, and SVM classifiers are used to differentiate between the spin images of nose tip and inner eye corners.

#### 4.4 Joint use of 2D and 3D

Since most of the 3D acquisition devices provide also 2D color-texture image over an identical sampling grid, the prospect of utilizing 3D and 2D data jointly becomes attractive. Such multi-modal approaches have already shown promising results [42,95]. For example, Boehnen and Russ [42] used range and color data jointly. First, range data is used to eliminate the background and to constrain the facial component search area. A YCbCr color model in conjunction with a geometric-based confidence measure is used to segment skin regions and determine eye and mouth regions. Geometry-based modality aids in the selection of the best landmark set, and it is based on the 3D measurements made over the range data.

Salah and Akarun [95] compared 3D-based landmark extraction with 3D-aided 2D landmark extraction. They model Gabor jet features statistically as a mixture of Gaussians in a lower-dimensional manifold, and to this end they use mixture of factor analyzers. They concluded that under favorable conditions (e.g., FRGC v.1) 2D and 3D systems perform on a par. Under unfavorable conditions (e.g., FRGC v.2), 3D performs better on nose tip and eye corners, though the detection rate is lower at mouth corners. However, under adverse conditions, the 2D and even 3D-assisted 2D algorithms completely fail. Open-mouthed facial expressions is one of main reasons for their lower localization performance, and they observed that the wrinkles between lower lip corner and chin also cause false positives for the mouth corners in 3D.

Akakin and Sankur had addressed the 3D landmarking problem as a combinatorial search [27]. Recently Sunko et al. [96] have managed the combinatorial problem using RANSAC algorithm. First, they find the reliable features



using spin images as features, and the missing ones are regressed using the multivariate Gaussian model encompassing all 3D landmark coordinates. To sort out the correct landmarks from the multitude of candidate points, they use all combinations of four points, and RANSAC is used as the basis of the feature matching procedure. The median of the closest candidates for the missing landmarks is considered. The cost function consists of a part accounting for the reconstruction error, that is the PCA-instrumented shape fitting term for the found landmarks, while the other part accounts for the distance from the inferred landmarks to their closest candidates.

3D face landmarking methods are summarized in Table 3.

## 5 Recent progress in face landmarking

Interest in face landmarking seems recently to be revamped as attested by the flurry of articles in the last five years. In contrast to the efforts of the last two decades, recent studies are characterized by the following: (i) A wider employment of machine learning techniques ranging from random ferns to the aggregation of weak learner outcomes to result in a more robust estimator; (ii) Confronting a wider range of out-of-plane face poses, notably yaw angles; (iii) Training and testing across several databases and increasing use of “faces in the wild”, e.g., real world faces in unconstrained environments collected mostly from the web; (iv) A more pronounced employment of regression techniques, whether support vector regression (SVR) or Random Forests, in lieu of classification techniques.

Below we discuss a sampling of algorithms introduced in recent articles, which we think are prototypical of recent progress in face landmarking. These articles could also have been discussed in Section 3 under the appropriate categories. However, we opt to discuss them separately to give a flavor of recent research trends in landmarking. Furthermore, we have added suggestive subtitles to indicate what we think to be the more innovative aspect of the work, but otherwise they do not denote a strict categorization.

### 5.1 Regression methods

The two-tier approach of Valstar et al. [70] uses in the first level surrounding image information to predict landmark location via support vector regression (SVR), and in the second level, the global shape information via a Markov Network. The regressor simplifies the landmark search in contrast to exhaustive sliding-window search with a template window. Briefly, they use Haar-like filters as descriptors of local appearance, benefiting also from the speed advantage of the integral images. The search for the initial seven fiducial landmarks exploits a prior model of landmark locations in the bounding box of the face. The

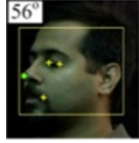

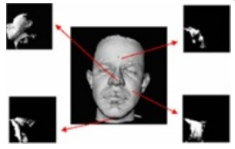
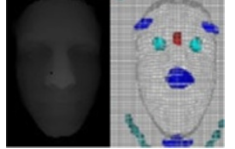


features are selected by the Adaboost regression, which uses multiridge regression as the weak classifier, and their ultimate number is determined not by Adaboost itself, but subsequent cross-validation using SVRs. Once the fiducial landmarks are consolidated, they generate hypotheses for the positions of the remaining 15 “unstable” landmarks, which are refined then with another application of SVR. A bank of regressors predict the distance and angle to the target landmark, and their votes are combined via the median operator. Finally, the global information is put into use by means of a Markov Network, which uses the learned spatial relationships between landmarks and penalizes improbable landmark configurations. Network nodes are not landmark locations per se, but relationships between pairs of landmarks, that is, vectors that point from one landmark to another. Since the angular differences and the length ratios of these vectors are used, planar rotation and scaling problems as well as any initial Viola-Jones face detector errors are automatically taken care of. The innovative aspects of this study is a relatively new way of combining local and global information, rightly called Boosted SVR coupled with Markov Networks: BORMAN. In a follow-up version of this work [47], they use evidence-driven sampling, and test the enhanced algorithm on a much larger set of conditions and of databases.

Cao et al. [97] point out that local evidence is sufficiently strong only for a few prominent landmarks, but otherwise most others are not salient enough and cannot be reliably characterized by their image appearance, and therefore shape constraint is essential. Their method is regression based where the shape constraint is realized in a nonparametric manner. Their nonparametric approach is based on the fact that the regressed shape is a linear combination of all training shapes. An interesting aspect is that instead of using the regressors in parallel and fusing their result as in [98] the authors use sequential regressors, where each one in the sequence uses the image information and the shape estimated from the previous stage of regression. Furthermore, the regressed shape is always constrained to reside in the linear subspace constructed by all training shapes. This guarantees the plausibility of the shape as well as global consistency.

### 5.2 Tree-structured search

Zhu and Ramanan [99] address the three linked problems of face detection, face pose estimation and face landmarking jointly. Since pose is part of estimation, the algorithm practically works as a multiview algorithm. In contrast to [47,70], where local and global information are invoked in succession, this algorithm is shape driven, and local and global information are merged right from beginning. This is implemented by considering several (30 to 60) local patches that are connected as a tree, which

**Table 3 An overview of the 3D face landmarking algorithms**

| Work  | Highlights of the method   | Domain knowledge used   | Landmark types   |
|---|--|---|--|
| Lu & Jain [88], 2006  | Shape index and cornerness information are fused into a field where extrema are searched at conjectured locations. Since face orientation is estimated, the method is robust against pose (i.e., yaw).   | Nose tip is detected as the peak of the central vertical profile. Prior location probability of the eye and mouth corners vis-à-vis the nose tip. Anthropometric distances between landmark points measured in world coordinate system form a constraint set. | Mouth, inner eye corners, nose and chin tip.<br>                            |
| Gökberk et al. [87], 2006                                   | An Average Face Model with 10 landmark points is aligned to the scene face via Iterative Closest Point algorithm. Initialized landmark positions are corrected via shape descriptors of Gaussian curvature, mean curvature, surface normals and landmark distances to the face symmetry plane. | 3D Average Face Model introduces both face geometry and local shape information.  | 10 landmarks ( <i>m7</i> plus philtrum <sup>a</sup> , nasion and chin).<br> |
| Conde & Cabello [94], 2006                                  | Mean curvature field of the face reveals the high curvature extrema; spin images at these extrema are classified via SVM as eye inner corners and nose tip.  | None.   | Endocanthion <sup>b</sup> and nose tip.<br>                                 |
| Akagunduz & Ulusoy [91], 2007                               | Mean and Gaussian curvatures are calculated at many scales, and organized as a space-scale Gaussian pyramid (UVS). Surface shape properties within the connected components in the UVS space are investigated as being eye pits, chin and nose protuberances.                                  | Topological graph to regularize the search is only suggested.   | Eye pits, nose tip and pit, chin.<br>                                      |
| Salah & Akarun [95], 2006 and Dibeklioglu et al. [90], 2008 | Gabor jets are statistically modeled as incremental mixture of factor analyzers (IMoFA) to generate a lower-dimensional manifold. IMoFA is run on the difference image of the Gaussian and mean curvature fields.  | Nose tip heuristics.  | <i>m7</i> landmark set.<br>   |
| Nair & Cavallaro [93], 2009                                 | PDM: Point Distribution Model, i.e., a parametrized model of the 49 3D landmark configurations is computed. The PDM is fitted to the face driven by local curvedness and shape index information.  | (i) PCA model of the 49 landmark points; (ii) face heuristics to prune out combinations of candidate landmarks to arrive to plausible shapes.   | 49 upper face landmarks.<br>  |

<sup>a</sup>Philtrum is the vertical groove between the base of the nose and the border of the upper lip.

<sup>b</sup>Endocanthion is the point at which the inner ends of the upper and lower eyelid meet.

collectively describe the landmark related region of the face; in other words, the patch-based face graph models the RoI of the detected face and incorporates its pose and landmark information. This approach is an adaptation of the idea of tree-structured pictorial structures [100]. In more detail, each patch is characterized by a HOG descriptor [33], and these patches are connected with quadratic springs in order to configure a shape. The authors employ a mixture of trees where each tree corresponds to a pose or to an expression in the frontal

pose. The final shape is determined by the maximum likelihood tree structure that best explains the landmark locations for the given mixture, assuming that the landmarks are Gaussian distributed. In effect, one infers the face pose and landmarks by maximizing over all mixtures and over all possible shapes given the patch HOGs. They also investigate the sharing of parts [101], and not surprisingly, non-shared model is better, albeit slightly, in both pose and landmark accuracies. Tree-structured pictorial structures have also been successfully applied

to face recognition by Everingham et al. [102], where the local appearance of each landmark are learned by a variation of Adaboost algorithm with Haar-like features [38]. Similarly, Uricar et al. [103], inspired by pictorial structures, jointly optimize appearance similarity and deformation cost with a parameterized scoring function where the parameters are learned from manually annotated instances using the structured output SVM classifier.

### 5.3 Random forests and ferns

Dantone et al. [104] propose pose-dependent landmark localization scheme that is achieved by conditional random forests. While regression forests try to learn the probability over the parameter space from all face images in the training set, conditional regression forests learn instead several conditional probabilities over the parameter space, and thus can deal with facial variations in appearance and shape. The head pose is quantized into five segments of “left profile, left, front, right and right profile” faces and specific random forests are trained. The local properties of a patch is described both by texture and by 2D displacement vectors that are defined from the centroid of each patch to the remaining ones. Specifically, texture is described by Gabor filter responses in addition to normalized gray values in order to cope with illumination changes. Training of conditional random forests is very similar to random forests; the main difference is that the probability of assigning a patch to a class is conditioned on the given head pose. This approach is able to deliver located landmarks in a query image at real-time speed.

Efraty et al. [105] contain a very extensive study on landmarking performance using a multitude of face databases. Their version of local and global information paradigm operates as follows: locally, unions of simpler polygonal sub-shapes represent groups of landmarks; globally, all sub-shapes are deformed in parallel toward their target landmark positions. The global search is instrumented via agglomerate fern regressors. The sub-shapes are triangles, whose vertices should eventually settle on the corresponding landmark positions. Since multiple instances of sub-shapes are initialized, the scheme is claimed to be robust against pose, illumination and expression artifacts. The final landmark positions are computed as the mean of these parallel instantiations, and their variance indicates the reliability of the landmark. In summary, while their preprocessing step and shape model are fairly standard, the originality of the method lies in the bank of regressors, which are charged with the duty of predicting the deformations for all instances of sub-shapes. The algorithm fuses every few iterations the predicted positions for the landmarks shared by several sub-shapes.

### 5.4 Bayesian approach

Belhumeur et al. [106] use innovatively a fully Bayesian approach to deduce landmark positions from local evidences. An interesting aspect of their work is that these evidences, that is, the local detector outputs are collected from a cohort of exemplars (sample faces with annotated landmarks), which thus provide non-parametrically the global model information. In other words, anatomical and geometrical constraints on facial landmarks are implicit in the exemplars. Depending upon the choice of the exemplars, localization robustness can be obtained against a large range of real-world variations in pose, expression, lighting, makeup and image quality. The local detector itself consists of a sliding window whose size is proportional to IOD and which collects SIFT features. The normalized SVM score of the SIFT feature set,  $d$ , gives local likelihood of a landmark at position  $x$ :  $P(x|d)$ . In the next stage, the global detector models the configurational information of the ensemble of fiducial points. Thus the joint probability of the locations of the  $n$  landmarks,  $\mathbf{X} = [x_1, \dots, x_n]$ , given the vector of their local detector outputs,  $\mathbf{D} = [d_1, \dots, d_n]$ , that is  $P(\mathbf{X}|\mathbf{D})$ , is maximized. It is interesting to note that this method surpasses in accuracy the performance of the manual landmarking in most of the 29 landmarks considered.

### 5.5 Semi-supervised learning

Tong et al. [107] address the tedious and often imperfect task of manual landmark labeling, and suggest a scheme to partly automate it. In their method, a negligible percentage (e.g., 3%) of faces need to be hand labeled, while the rest of the faces are automatically marked. This is realized by propagating the landmarking information of the few exemplars to the whole set. The learning is based on the minimization of the pairwise pixel differences resulting in two error terms: The penalty in one term controls the warping of each un-marked image toward all other un-marked images, so that they become more alike irrespective of the content. The penalty in the other term controls the warping of un-marked images toward marked images, and it is here that the physical meaning of the content is imposed. The warping function itself can be a global affine warp for the whole face, or a piecewise affine warp to model a non-rigid transformation.

### 5.6 Multi-kernel SVM

Rapp et al. [108] start with the two major patches on the face: one covering the eye region and the other roughly the mouth region. For testing, pixels in the respective regions to be part of a target landmark; texture data is extracted using the multiresolution windows (progressively smaller nested windows) that capture information ranging from global to local view. The pyramidal information is not concatenated; instead, every resolution level

is fed into a different kernel and the convex combination of these kernels, each dedicated to a resolution level, forms a multi-kernel SVM. The SVM is trained using center-surround architecture, with the surround windows forming the negative examples. Following the discovery of landmark points, initially without any spatial relationship among them, point distribution models are invoked to reach to plausible shapes. The point distribution models are particularized to the eye-eyebrow pairs and to the mouth. The shape alternatives are evaluated using Gaussian mixture models (GMMs), so that the point combinations that possess the highest sum of SVM scores and that fit best to the learned models are selected.

### 5.7 Extended template

Kozakaya et al. [98] solve the landmarking problem using multiple voting of extended templates. An extended template is defined for each point as a combination of three parts, of a local descriptor at the sampling point, of its directional vector pointing to the given landmark position, and thirdly, of the local likelihood pattern around the landmark positions. Sampling points are taken on a regular grid. Every local descriptor has  $N$  vectors pointing to the  $N$  landmark points and the associated local likelihood patterns, where local likelihoods are obtained from the HOG vectors. In other words, local information resides in the HOG features of the two, sampling and target landmark localities while the global shape information resides in the joint treatment of the  $N$  landmarks. This scheme attains robustness by two means, by the large number of sample points around each target on the face and by nonlinear fusion of the resulting pointing vectors. In the fusion stage, pointing vectors are weighted inversely proportional to the local appearance error and are combined robustly with least median of the squares.

In summary, from the above articles and from the many others reviewed but not accommodated in the text, we have observed the leitmotiv of coordinating and exploiting the local image evidence and the global shape information at various levels of sophistication.

## 6 Experimental study and comparisons

### 6.1 Standard databases

Since standard databases are compulsory for experimental assessment and performance comparisons of algorithms, we briefly review the most relevant face databases suitable for landmarking studies. Ground truth data, i.e., manually landmarked spatial positions of the landmark points is much desired for referenced landmarking performance. One way to obtain the ground-truth data is to employ manual work or use Amazon mechanical turk (MTurk) scheme to carry out this tedious task. Typically, several folds of independent manual landmarking are run since there is rarely full agreement between the markers, and

the ground truth positions are taken as the mean of these folds (typically three). An interesting result is reported in [106] where the automatic landmark detector proves to be more consistent than three human annotators, especially for eyebrows and chin tip. We would like to note that quality of the manual landmarking is critical since higher consistency can boost performance. A case in point is the annotation with Mechanical Turk where the quality checks are loose; instead the average variance of trained annotators tends to be much lower.

We present prominent ground-truthed face databases according to two categories: databases under controlled conditions and databases without any control and conditioning. Controlled databases are collected within the framework of a defined experimental setup using one or more of the four control instructions: (i) different facial expressions; (ii) occlusions; (iii) head rotations; and (iv) illumination variations. For example, CMU Multi-PIE database [109] is a good testbed for rotations, and the Bosphorus database [110] is very rich in the variety of facial action units and facial expressions. Uncontrolled databases, on the other hand, are collected without any directives given to the subjects, and they are appropriately called sometimes “faces in the wild”. Recently, databases culled from social network sites such as google.com, flickr.com, facebook.com have stirred a lot of interest, first, because they provide more realistic and challenging databases, and second, due to the huge potential of web sources in sharp contrast to the laborious process of building controlled databases. The downside is, of course, uncontrolled faces are seldom labeled.

#### 6.1.1 Controlled databases

The more commonly used controlled face databases are as follows:

**Aleix-Robert face database (AR'98):** AR face database contains 4000 color images of 126 people [111]. There are strict constraints on the pose of subjects. Each subject has 13 different images, including 4 facial expressions, 3 illumination variations, 2 occlusions (wearing sun glasses and scarf) and 4 occlusions on top of illumination changes (e.g., wearing sun glasses and leftward illumination); these images are captured in two sessions with a two-week interval. Ding and Martinez[68] have provided also 130 manually annotated landmarks on the contour of faces and of the facial components. The dataset is available on request at: <http://www2.ece.ohio-state.edu/~aleix/ARdatabase>.

**Extended M2VTS database (XM2VTS'99):** XM2VTS database contains video recordings of 295 people [112]. All recordings are collected in four

sessions over a period of four months. Each recording is captured while the subject is speaking or rotating his head along yaw and pitch directions as instructed. The collection consists of 2360 color images, sound files and 3D face models. Note that 30% of the images are of poor quality due to motion blur on faces or due to closed eyes and hair occlusion on face. The data set is available on request at: <http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb>.

**Cohn-Kanade database (CK'00):** The 97 subjects in the first released portion of the Cohn-Kanade AU-Coded Facial Expression Database [113] were instructed by an experimenter to perform a series of facial displays that include single action units and combinations of action units. Totally, 486 video sequences were recorded where each sequence begins with a neutral expression and leads to the apex of the target expression. The extended version CK+'10 [114] contains 43 FACS coded action units and 8 studied facial expressions as well as spontaneous smile expressions. Faces in all sequences are frontal and free of illumination variations. In this release, the number of subjects is further increased to 123, resulting in totally 593 video sequences. For each sequence, a keyframe is manually annotated, and in the remaining frames annotated landmarks are automatically aligned to face by the gradient descent AAM fitting algorithm in [115]. The dataset is available for distribution at: <http://www.pitt.edu/~jeffcohn/CKandCK+.htm>.

**Bosphorus 3D face database (Bosp'08):** This database includes 2D and 3D facial images of 105 subjects imaged with a great variety of action units (more than 35 per person), in addition to the six universal motions and several occlusion instances. The head poses of subjects can have mild rotation, yaw and/or tilt, but otherwise faces are free of the illumination artifacts [110]. Since faces are captured both in 2D and in 3D with a structured light camera, the two modalities are registered. The faces annotated with 24 landmarks in addition to being FACS (Facial Action Coding System) [116] coded, therefore this database can be used as a testbed for 3D landmarking as well as for comparative study of 2D and 3D landmarking. Information on how to obtain the dataset can be found at: <http://bosphorus.ee.boun.edu.tr>.

**CMU multi-pose, illumination, and expression face database (Multi-PIE'08):** The CMU PIE face database [109] contains over 750,000 facial images of 337 people. People are imaged across 15 different poses, under 19 different illumination conditions, and with 6 different expressions. The number of annotated landmarks, provided only for a small

subset of the dataset, varies between 39 and 68 since all landmarks are not visible in profile faces. Details on obtaining the dataset can be found at: <http://www.multipie.org>.

While the collections described above are the more important controlled databases for landmarking, we find it worth mentioning the following: FRGC data set of 2D and 3D face images [26], FERET database containing face rotations from frontal to left/right profiles [117], MMI Facial Expression database [118], UHDB11 database containing rotations and light variations [119] etc.

### 6.1.2 Uncontrolled databases

Uncontrolled face databases are becoming increasingly popular in face detection, recognition, pose estimation and in landmarking studies as they are more realistic and challenging. The following datasets are of interest:

**BioID face database (BioID'01):** BioID database is one of the most popular benchmarks for landmarking algorithms. The database consists of 1521 gray level images [120], collected within the framework of FGNet project, European Working Group on face and gesture recognition. These images show the frontal views of faces of 23 different subjects, which are recorded during several sessions in uncontrolled conditions using a web camera within an office environment. Compared to the controlled databases, this dataset features a larger variety of illumination conditions, backgrounds and face sizes.

Faces are manually annotated on 20 landmarks for the purposes of facial analysis and gesture recognition. The data is publicly available at: <http://www.bioid.com/downloads/software/bioid-face-database>.

### Labeled face parts in the wild database

**(LFPW'11):** LFPW database contains 3000 face images downloaded from the web [106]. Face images are automatically detected by a commercial face detection system. This database exhibits a large variety in appearance, e.g., facial expressions, pose, age, ethnicity, imaging and environmental conditions etc., and also includes manipulated photos, cropped faces from movie scenes with extreme make-up and clothing. However, the face detector fails to detect near- and in-profile faces, which therefore are excluded from the database. There are 29 manually landmarked points, and this number can go up to 35 whenever landmark points on the ear are visible. Manual annotations are obtained by employing three MTurk workers and the ground truth is determined by averaging the three manual annotated locations. A subset of the database is made available and divided



into two protocols, i.e., training and testing, at: <http://www.kbvt.com/LFPW/>.

**Annotated facial landmarks in the wild**

**(AFLW'11):** In the same vein, AFLW, contains real-world images collected with a large variety in appearance from Flickr [121].

This database is an order of magnitude bigger and more complex as compared to LFPW dataset in that the number of images and with annotated face landmarks totals to 25,993 and it possesses a much larger variety of face poses (the ratio of the non-frontal faces is 66% including up to  $\pm 90^\circ$  head rotations). Annotated locations are provided for 21 landmark points in addition to the ellipses and rectangles enclosing the face can be found. The database is available on request at: <http://lrs.icg.tugraz.at/research/aflw>.

**Annotated faces in the wild (AFW'12):** AFW dataset [99] is another example of in-the-wild collections and differs from the previous datasets [106,121] in that it consist of 205 images including

more than one face per image, 468 faces in total. This renders AFW relatively more challenging as it contains images with highly cluttered background and with large variations both in face scale and pose, i.e., it is possible to observe frontal and non-frontal faces or close up and distance shots in a single image. Each face is labeled with a bounding box, 6 landmark locations and head rotation angles. The database is not available online yet, but more information can be found at: <http://www.ics.uci.edu/~xzhu/face>.

Another potentially useful database is called labeled faces in the wild (LFW'07) [122], which contains 13,233 facial images downloaded from the web. There are 5749 distinct subjects, 1680 of which have more than one image. Manual landmark annotations are not provided for this dataset, though recently Dantone et al. [104] had MTurk workers to annotate the locations of 10 landmark points.

The main characteristics of the aforementioned datasets are summarized in Table 4.

**Table 4 Overview of databases for landmarking studies**

|                    | Number of subjects | Number of images | Number of landmarks* | Modality             | Control tag**        | Reference works  |
|--------------------|--------------------|------------------|----------------------|----------------------|----------------------|--|
| Controlled         |                    |                  |                      |                      |                      |  |
| AR'98 [111]        | 126                | 4000             | 130                  | color image          | e, i, o, f           | [31,123]<br>[28,108]                                     |
| XM2VTS'99 [112]    | 295                | 2360             | 68                   | video sequence       | o, f, nf, b          | [59,80,81]<br>[41,47,85]<br>[31,83,124]<br>[28]          |
| CK+'10 [114]       | 123                | 10,734           | 68                   | video sequence       | e, f                 | [28,44,108]  |
| Bosp'08 [110]      | 105                | 4666             | 22–24                | color, 3D data       | e, o, f, nf, r       | [28,90]  |
| Multi-PIE'08 [109] | 337                | 750,000          | 39–68                | color image          | e, i, f, nf, r       | [47,105,123]<br>[85,99,108]                              |
| Uncontrolled       |                    |                  |                      |                      |                      |  |
| BioID'01 [120]     | 23                 | 1521             | 20                   | gray-scale image     | e, i, o, f, nf, s    | [81,82,105]<br>[47,70,83]<br>[97,106,124]<br>[28,31,108] |
| Uncontrolled       |                    |                  |                      |                      |                      |  |
| LFPW'11 [106]      | –                  | 3000             | 29–35                | color,<br>gray-scale | e, i, o, f, nf, s    | [97]   |
| AFLW'11 [121]      | –                  | 25,993           | 21                   | color                | e, i, o, f, nf, r, s | –  |
| AFW'12 [99]        | 468                | 205              | 6                    | color                | e, i, o, f, nf, r, s | –  |

\*Number of landmarks can vary upon visibility, e.g., 68 landmarks points are not available for profile faces.

\*\* Acronyms of the control tags are as follows: i: illumination changes, f: facial expression, o: occlusion, r: head rotation, f: frontal face, nf: near-frontal (up to 15° yaw), s: scale change, b: blur. The rightmost column lists representative methods that have used the corresponding database.

**Table 5 Experimented face landmarking algorithms**

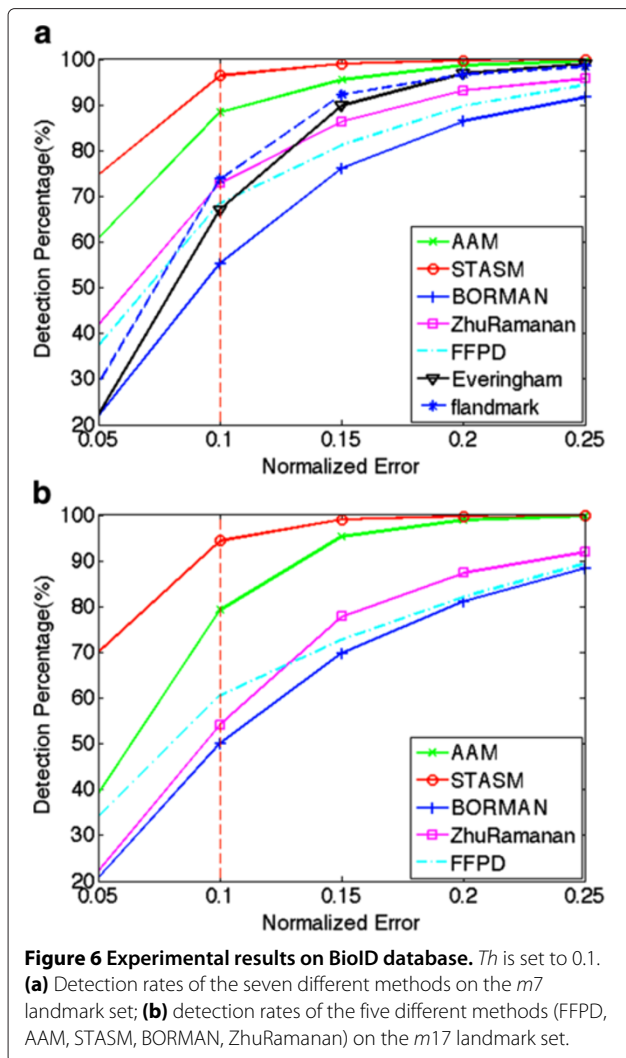
| Acronym                      | Face detector                                 | #Landmarks | Training set      | Face pose and expression       | Processing time per image* |
|------------------------------|---|------------|-------------------|--------------------------------|----------------------------|
| FFPD [44]                    | Haar feature based<br>GentleBoost classifier  | 20         | CK frontal faces  | Frontal; Neutral               | 0.85                       |
| AAM [125]                    | Viola-Jones face<br>detector                  | 66         | Multi-PIE, XM2VTS | Near-frontal;<br>Expression    | 0.12 s                     |
| STASM [31]                   | Viola-Jones and Rowley<br>face detector [126] | 76         | XM2VTS, AR        | Near-frontal;<br>Expression    | 0.18 s                     |
| BORMAN [70]                  | Viola-Jones face detector<br>detector         | 22         | FERET, MMI        | Near-frontal;<br>Expression    | 65 s                       |
| ZhuRamanan <sup>†</sup> [99] | A mixture of tree<br>structured part models   | 68         | Multi-PIE         | Free of pose and<br>expression | 25 s                       |
| Everingham [102]             | Viola-Jones face detector                     | 9          | Consumer images   | Near frontal                   | 0.4 s                      |
| flandmark [103]              | Description NA                                | 7          | LFW               | Near frontal;<br>expression    | 0.12 s                     |

\*Average run time on BioID database with a CPU of 2.50 GHz and 8 GB RAM. Each image has a resolution of  $384 \times 286$ .

<sup>†</sup>Trained model with 1050 parts.



**Figure 5 Results of the tested algorithms.** The columns refer to different algorithms from left to right: FFPD, AAM, STASM, BORMAN, ZhuRamanan, Everingham, and flandmark. The rows correspond to three different datasets: BioID (upper row), CK+ neutral (middle) and CK+ expressions (lower row).



### 6.2 Comparative assessment of landmarking methods

In this section, we present the performance of selected landmarking algorithms comparatively as tested on three diverse face databases. Among several candidates, we selected seven landmarking algorithms as listed in Table 5. The rationale of selection was that the algorithm was deemed to be representative of the state of the art and

for which an open software was available. For a thorough comparison, we ran the algorithms both on controlled databases (CK+ and Bosphorus) and an uncontrolled database (BioID). With the controlled databases, we aim to study in detail the influence of two confounding factors: facial expression (CK+) and head pose (Bosphorus). These tested databases have been particularly selected since none of the algorithms listed in Table 5 have been trained on these three databases. We believe that this point is important for a fair evaluation of algorithms. Face literature ranging from face recognition to pose estimation reports that testing and training within the same database, even with non-overlapping training and testing subsets, can yield optimistic results as compared to experiments where databases used for training and testing are different. The experiments on the Bosphorus database is limited, because the considered methods were trained only with frontal and near-frontal faces, and this did not permit us to make a fair comparison against head poses on Bosphorus database.

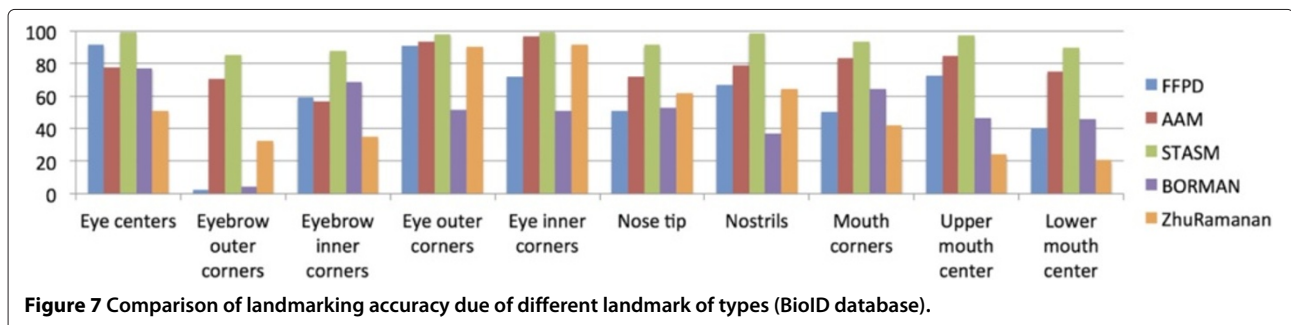
The results of the tested algorithms are illustrated for BioID and CK+ database in Figure 5. The detailed experimental results are presented in the sequel.

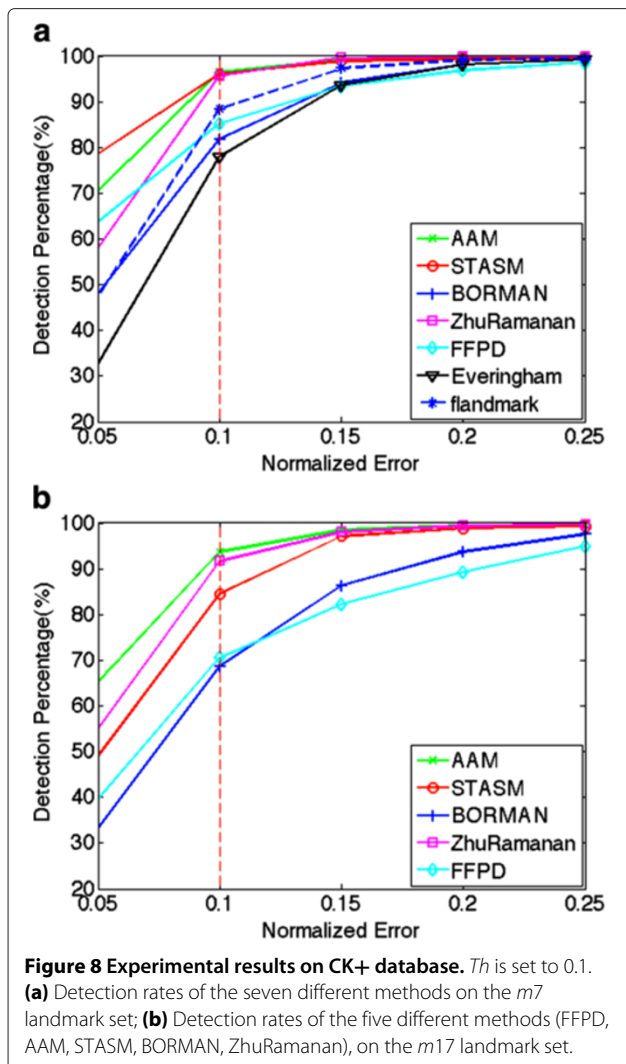
Recall that there are two subsets of landmarks that are often used for testing, and these are referred to as the  $m7$  and  $m17$  subsets:

- $m7$  set: This set consists of four eye, one nose and two mouth landmarks. According to Figure 2, these are 28, 30, 32, 34, 41, 46, and 52.
- $m17$  set: This set consists of four eyebrow, six eye, three nose, and four mouth landmarks. These are 16, 19, 22, 25, 28, 30, 32, 34, 40, 41, 42, 46, 49, 52, 55, 63, and 64 as illustrated in Figure 2.

### 6.3 Tests on BioID database

We first compared the seven different methods on the BioID dataset, which is the most frequently used uncontrolled testbed for landmarking algorithms. Figure 6a reports the performance results for the  $m7$  set and Figure 6b reports those for the  $m17$  set. Since the two methods, Everingham and flandmark are not designed to





output  $m17$  landmarks, there are only five competitors for the  $m17$  set: FFPD, AAM, STASM, BORMAN, and ZhuRamanan. We observe that only STASM qualifies and has the best performance among all, reaching above 90% detection rate at the  $0.1 \times \text{IOD}$  threshold. AAM can be considered as a runner up for  $m7$  with a performance high in the upper 80%. One would have expected, for example, ZhuRamanan to achieve better performance over 90% on this uncontrolled database. One explanation would be that ZhuRamanan was trained with Multi-PIE database, a controlled database, while BioID is uncontrolled.

Recall that the  $m17$  set contains somewhat more difficult landmarks, such as mouth centers and four on the eyebrows, while  $m7$  contains the most fiducial ones. Hence it is natural that the  $m17$  curves are below those of  $m7$ . In Figure 7, a more detailed comparison based on landmark types is given. The figures reported in the bar chart are the average of the left

and right landmarks, e.g., “eye centers” means the average of the eye center accuracies of the left and right eye. We observed that STASM does overall quite well while the other remaining four methods have variable performance, some dramatically falling down on certain landmarks. We found out that FFPD and BORMAN detect almost the middle of eyebrows instead of the eyebrow outer corners. This is illustrated in the first and fourth columns of Figure 5. This explains the significant lower performance of eyebrow outer corner detection in Figure 7. However, even excluding eyebrow outer corners and evaluating the performance over the remaining 15 landmark points did not change the performance: STASM, AAM, FFPD, BORMAN, and ZhuRamanan.

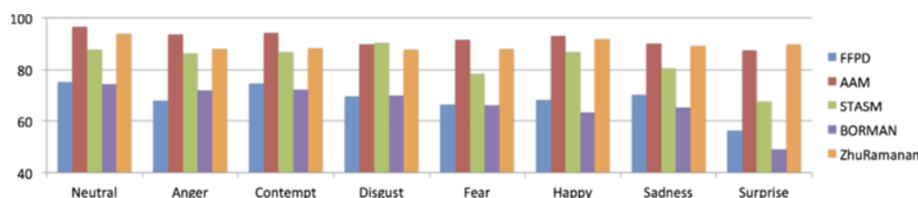
#### 6.4 Tests on CK+ database

We repeated the same experiments on the CK+ database, a controlled database, where we compared the seven and the five selected algorithms for  $m7$  and  $m17$  landmark sets, respectively. Figure 8a gives the performance of the seven algorithms on  $m7$  and Figure 8b of the five algorithms on  $m17$ . We notice that in this controlled database, AAM and ZhuRamanan algorithms surpass the performance of STASM, partly because the former two were trained on similar (but not identical) databases. Following the same reasoning, one can state that the FFPD algorithm has a low performance for expression faces since it is only trained with neutral faces. In Figure 8b, while the performances of STASM, FFPD, and BORMAN are significantly dropped, AAM and ZhuRamanan are much more robust under different facial expressions. Figure 9 reveals performance details per expression type. Again we see that AAM and ZhuRamanan are the best performers, uniformly over all expressions while STASM is a lagging runner-up.

A summary of experimental results as exceedance percentage of the  $0.1 \times \text{IOD}$  threshold is given on BioID and CK+ databases in Table 6. It is difficult to reach a fair and general conclusion on the methods because their performance depends critically on their training set. However, taking into account the computational efficiency and overall performance (please refer to Tables 5 and 6, respectively), AAM and STASM seem to be most promising for real time applications, while ZhuRamanan is the method to pursue for offline applications. It is true that the algorithm of ZhuRamanan is two orders of magnitude slower than, for example the AAM method. On the other hand, it is one of the best performing methods, especially in adverse conditions where most other methods fail.

#### 6.5 Tests on Bosphorus database

The Bosphorus database is rich in facial action units, in expressions and poses. Since expressions were addressed



**Figure 9** Comparison of landmarking accuracy for different facial expressions (CK+ database).

in the CK+ experiments, we used a subset of the Bosphorus database with yaw movement, that is, faces with yaw angles of 10°, 20°, 30°, 45°, and 90°. However, all algorithms fail since their code uses the Viola-Jones detector, which can only handle faces up to ±15° yaw angles. The only surviving algorithm is ZhuRamanan algorithm which does not explicitly depend upon the Viola-Jones face detector. For this reason, and partly because one cannot express performance figures for occluded landmarks, we limited ourselves to give illustrations of landmark detections for various yaw angles. As can be seen from Figure 10, ZhuRamanan can handle yaw rotations even if trained on a different database, which is another merit of this algorithm.

## 7 Conclusion

After surveying of face landmarking techniques, of recent research trends and comparative performance figures, we could draw the following conclusions:

- **State-of-the-art:** The successful methods are the model-based ones, which integrate landmark evidences from local patches with a global shape constraint. The two-tier approaches are methods that in the first tier extract fiducial landmarks, and in the second tier predict and consolidate landmarks with less informative features under the guidance of a face shape model are more successful. The coordination

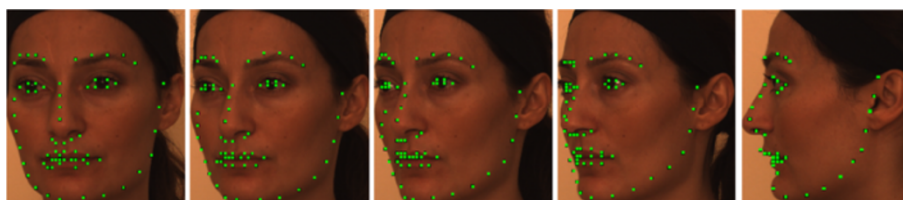
**Table 6** The detection accuracies of the algorithms on BioID and CK+ databases for the threshold set at 10% IOD value

|            | BioID database |     | CK+ database |     |
|------------|----------------|-----|--------------|-----|
|            | m7             | m17 | m7           | m17 |
| FFPD       | 65             | 60  | 83           | 69  |
| AAM        | 84             | 79  | 95           | 92  |
| STASM      | 93             | 94  | 95           | 83  |
| BORMAN     | 53             | 50  | 79           | 66  |
| ZhuRamanan | 69             | 51  | 94           | 90  |
| Everingham | 59             | NA  | 76           | NA  |
| flandmark  | 69             | NA  | 86           | NA  |
| Average    | 70             | 66  | 85           | 78  |

of the local and global information, or first- and second-tier operations is realized with a diversity of methods ranging from Bayesian prediction to SVR. It appears that the performance of algorithms in the last five years have improved to a point where for the *m17*, it is on a par with manual landmarking. In fact, if we limit our observations to the published results in the articles this signifies a few percentage points, like 2–3% of IOD. Outside the *m17* set, the accuracy remains within 5–8%. However, our experiments on the seven most prominent landmarking algorithms have revealed that these results are not always reproducible. More specifically, when testing and training database pairs are different than the ones mentioned in the article, there can be significant deviations from the announced results.

- **Landmarking under realistic and adverse conditions:** Face landmarking methods, being often local in nature, can be made more robust to intrinsic variability and acquisition conditions. It is possible to state that illumination effects can be mostly compensated by such preprocessing steps as LoG: Laplacian of Gaussian filtering or histogram equalization. Similarly, facial expressions and modest pose variations can be made up for by a richer set of training instances. The bane of landmarking remain severe pose variations, i.e., beyond 20° yaw angles and tilts, especially when self-occlusions occur. We assume that in-plane rotations can be corrected after detection of the face and of the eyes. It appears that hybrid methods like appearance-assisted geometry-based methods [9,58], 3D-assisted methods [88,90] or a connected battery of local templates as in [99] hold a good promise for success.
- **Robustness and ground-truthed databases:** The performances of algorithms may differ strongly from database to database. In fact, the across-database performance of the early algorithms, when they were trained on one database and tested on another database, showed this weakness, incurring sharp drops in performance. It is encouraging to witness that recent algorithms, notably [47,97,99,106,108] have robust performance across a number of different databases. The experimental results, though





**Figure 10** Results of the ZhuRamanan algorithm on Bosphorus database. Sample landmarked faces for different yaw rotations (from left to right): 10°, 20°, 30°, 45°, and 90°.

quite extensive in intent, has not yet revealed the ultimate and most fair comparison, in that methods have not been given chances to be trained on arbitrary combinations of databases. In fact, for a fair comparison, we suggest that methods should be tested in the LODBO manner: Leave One Database Out style, where algorithms are trained with all databases except one and then tested on the excluded one. Finally, this survey of methodological comparisons and the landmark databases should be extended to dynamic scenes to evaluate the concomitant problem of landmark tracking algorithms [35]. In fact, the landmark tracking problem itself deserve a separate review effort.

- **Methods to be explored:** We believe some of the promising research paths in landmarking techniques are the following: (i) Sparse dictionaries: The paradigm of recognition under sparsity constraint and building of discriminatory dictionaries seems one viable method. The discriminative sparse dictionary can be constructed per landmark [127,128] or collectively as in [129]; (ii) Adaboost selected features for multiview landmarking: Gabor or Haar wavelet features selected via modified Adaboost scheme where commonality and geometric configuration of landmark appearances is exploited [101]; (iii) Multiframe landmarking: Determination of landmark positions exploits the information in subsequent frames of a video, using, for example, a spatio-temporal representations [130,131].
- **Facial expression and gesture data mining:** Presently Internet contains at least 200,000 face videos [132], usually annotated with contextual information, and this number is rapidly increasing. This wealth of data provides an interesting opportunity to explore human facial expressions, in a sense, to data mine expressions across cultures, genders, ages and contexts. This source of face data is important because it has been pointed out that the lack of naturalistic, spontaneous expression data was a major roadblock in computer analysis of facial expressions. It has been pointed out that role-playing

expressions, that is facial expressions acted out as prompted by a controller differ in their dynamics and variety as compared to spontaneous expression of the same emotions. We believe robust landmarking will be instrumental for tapping this very rich web source of genuine human expressions.

In conclusion, facial landmarking has come a long way from its meager beginning at the end of eighties. The problem can be considered to be solved for near frontal faces with neutral to mild expressions, and adequate resolution. It appears that some of the successful algorithms can be run at video rates. On the other hand, for uncontrolled conditions involving arbitrary poses and expressions, the problem cannot yet be considered as thoroughly solved. Recent research results, however, give us a positive outlook.

#### Endnote

<sup>a</sup>Nasion is a distinctly depressed area directly between the eyes, just superior to the bridge of the nose.

#### Competing interests

The authors declare that they have no competing interests.

#### Acknowledgements

This study was supported by Boğaziçi University Scientific Research Projects (BAP) under Project No. 6533.

Received: 7 September 2011 Accepted: 15 January 2013

Published: 7 March 2013

#### References

1. Y Cohn, J Tian, T Kanade, Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(2), 97–114 (2001)
2. M Pantic, LJM Rothkrantz, Automatic analysis of facial expressions: the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(12), 1424–1445 (2000)
3. K Liu, A Weissenfeld, J Ostermann, X Luo, in *Proc. of Int. Conf. on Multimedia and Expo. Robust AAM building for morphing in an image-based facial animation system* (Hannover, Germany, 2008), pp. 933–936
4. S Ioannou, G Caridakis, K Karpouzis, S Kollias, Robust feature detection for facial expression recognition. *J. Image Video Process.* **2007**(2), 5–5 (2007)
5. U Park, AK Jain, in *Proc. of Int. Workshop on Video Processing for Security. 3D face reconstruction from stereo images* (Quebec City, Canada, 2006), p. 41
6. AA Salah, N Alyüz, L Akarun, Registration of 3D face scans with average face models. *J. Electron. Imag.* **17**(1), 011006 (2008)

7. N Pears, T Heseltine, M Romero, From 3D point clouds to pose-normalised depth maps. *Int. J. Comput. Vis.* **89**(2), 152–176 (2010)
8. A Lanitis, CJ Taylor, TF Cootes, Automatic face identification system using flexible appearance models. *Image Vis. Comput.* **13**(5), 393–401 (1995)
9. L Wiskott, JM Fellous, N Kruger, C von der Malsburg, Face recognition by elastic bunch graph. *IEEE Trans. Pattern Anal. Mach. Intell.* **7**, 775–779 (1997)
10. P Campadelli, R Lanzarotti, C Savazzi, in *Proc. of Int. Conf. on Image Analysis and Processing*. A feature-based face recognition system (Mantova, Italy, 2003), pp. 68–73
11. F Dornaika, F Davoine, in *Proc. of Int. Conf. on Pattern Recognition*, vol. 3. Online appearance-based face and facial feature tracking (Washington, DC, USA, 2004), pp. 814–817
12. J Cohn, A Zlochower, JJJ Lien, T Kanade, in *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*. Feature-point tracking by optical flow discriminates subtle differences in facial expression (Nara, Japan, 1998), pp. 396–401
13. H Çınar Akakin, B Sankur, in *Proc. of Joint Cost 2101 & 2102 Int. Conf. on Biometric ID Management and Multimodal Communication*. Analysis of head and facial gestures using facial landmark trajectories (Madrid, Spain, 2009), pp. 105–113
14. B Parkinson, Do facial movements express emotions or communicate motives? *Pers. Soc. Psychol. Rev.* **9**, 278–311 (2005)
15. P Ekman, WV Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. (Consulting Psychologists Press, Palo Alto, 1978)
16. HÇ Akakin, B Sankur, Robust classification of face and head gestures in video. *Image Video Comput.* **29**, 470–483 (2011)
17. J Bailenson, E Pontikakis, I Mauss, J Gross, M Jabon, C Hutcherson, C Nass, O John, Real-time classification of evoked emotions using facial feature tracking and physiological responses. *Int. J. Hum.-Comput. Stud.* **66**(5), 303–317 (2008)
18. A Pentland, B Moghaddam, T Starner, in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*. View-based and Modular Eigenspaces for Face Recognition (Seattle, Washington, 1994), pp. 84–91
19. B Heisele, P Ho, T Poggio, in *Proc. of IEEE Int. Conf. on Computer Vision*. Face recognition with support vector machines: global versus component-based approaches (Vancouver, British Columbia, Canada, 2001), pp. 688–694
20. J Shi, A Samal, D Marx, How effective are landmarks and their geometry for face recognition? *Comput. Vis. Image Understand.* **102**, 117–133 (2006)
21. Y Tong, Q Ji, in *Proc. of Int. Conf. on Pattern Recognition*, vol. 1. Multiview facial feature tracking with a multi-modal probabilistic model (Washington, DC, USA, 2006), pp. 307–310
22. J Chen, B Tiddeman, in *Proc. of IEEE Int. Conf. on Image Processing*. Robust facial feature tracking under various illuminations (Atlanta, GA, USA, 2006), pp. 2829–2832
23. RPW J Wiegardt, C von der Malsburg, in *Proc. of European Conference on Computer Vision*. Gabor-based feature point tracking with automatically learned constraints (Copenhagen, 2002)
24. L Teijeiro-Mosquera, JL Alba-Castro, Performance of active appearance model-based pose-robust face recognition. *Comput. Vis.* **5**(6), 348–357 (2011)
25. Y Hu, M Zhou, Z Wu, A dense point-to-point alignment method for realistic 3D face morphing and animation. *Int. J. Comput. Games Technol.* **2009**, 9 (2009)
26. PJ Phillips, PJ Flynn, T Scruggs, KW Bowyer, J Chang, K Hoffman, J Marques, J Min, W Worek, in *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, vol. 1. Overview of the face recognition grand challenge, (2005), pp. 947–954
27. HÇ Akakin, B Sankur, in *Proc. of Conf. on 3DTV*. Robust 2D/3D face landmarking (Kos, Greece, 2007), pp. 1–4
28. H Dibeklioglu, AA Salah, T Gevers, A statistical method for 2-D facial landmarking. *IEEE Trans. Image Process.* **21**(2), 844–858 (2012)
29. KE Ko, KB Sim, in *Proc. of Int. Symposium on Industrial Electronics*. Development of advanced active appearance model for facial emotion recognition (Seoul, Korea, 2009), pp. 1019–1022
30. A Asthana, A Khwaja, R Goecke, in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. Automatic frontal face annotation and AAM building for arbitrary expressions from a single frontal image only (Cairo, Egypt, 2009), pp. 2445–2448
31. S Milborrow, F Nicolls, in *Proc. of European Conf. on Computer Vision*. Locating facial features with an extended active shape model (Marseille, France, 2008), pp. 504–513
32. DG Lowe, Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
33. N Dalal, B Triggs, in *Proc. of Conf. on Computer Vision and Pattern Recognition*, vol. 1. Histograms of oriented gradients for human detection (San Diego, CA, USA, 2005), pp. 886–893
34. K Seshadri, M Savvides, An analysis of the sensitivity of active shape models to initialization when applied to automatic facial landmarking. *IEEE Trans. Inf. Forensics Secur.* **7**, 1255–1269 (2012)
35. Y Tie, L Guan, Automatic landmark point detection and tracking for human facial expressions. *J. Image Video Process.* **2013**, 8 (2013)
36. E Meyers, L Wolf, Using biologically inspired features for face processing. *Int. J. Comput. Vis.* **76**, 93–104 (2008)
37. X Xie, KM Lam, An efficient illumination normalization method for face recognition. *Pattern Recogn. Lett.* **27**(6), 609–617 (2006)
38. P Viola, M Jones, Robust real-time object detection. *Int. J. Comput. Vis.* **57**(2), 137–154 (2001)
39. FY Shih, CF Chuang, Automatic extraction of head and face boundaries and facial features. *Special Issue on Inf. Comput. Sci. Intell. Syst. Appl.* **158**, 117–130 (2004)
40. S Başkan, MM Bulut, V Atalay, Projection-based method for segmentation of human faces and its evaluation. *Pattern Recogn. Lett.* **23**(14), 1623–1629 (2002)
41. S Arca, P Campadelli, R Lanzarotti, A face recognition system based on automatically determined facial fiducial points. *Pattern Recogn.* **39**, 432–443 (2006)
42. C Boehnen, T Russ, in *Application of Computer Vision, 2005. WACV/MOTIONS '05*, vol. 1. A fast multi-modal approach to facial feature detection (Breckenridge, CO, USA 2005), pp. 135–142. doi:10.1109/ACVMOT.2005.5
43. A Gunduz, H Krim, in *Facial feature extraction using topological methods, 1-I-673-6*. *Proc. of Int. Conf. on Image Processing*, (2003), pp. 14–17. doi:10.1109/ICIP.2003.1247051
44. D Vukadinovic, M Pantic, in *Proc. of IEEE Int. Conf. on Systems, Man and Cybernetics*, vol. 2. Fully automatic facial feature point detection using gabor feature based boosted classifiers (Hawaii, 2005), pp. 1692–1698
45. W Hizem, Y Ni, B Dorizzi, Near infrared sensing and associated landmark detection for face recognition. *J. Electron. Imag.* **17**, 11005 (2008)
46. S Tsekeridou, I Pitas, in *Proc. of European Signal Processing Conf.*, vol. 1. Facial feature extraction in frontal views using biometric analogies (Island of Rhodes, 1998), pp. 315–318
47. B Martinez, MF Valstar, X Binefa, M Pantic, Local evidence aggregation for regression based facial point detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **99**, 1 (2012)
48. R Hsu, AK Jain, Face detection in color images. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(5), 696–706 (2002)
49. LZ Fang, YZ Sheng, AK Jain, WY Qiong, in *Proc. of Int. Conf. on Computational Intelligence and Multimedia Applications*. Face detection and facial feature extraction in color image (Xi'an, China, 2003), pp. 126–130
50. KH Seo, W Kim, C Oh, JJ Lee, in *Industrial Electronics, ISIE 2002. Proceedings of the 2002 IEEE International Symposium on*, vol. 2. Face detection and facial feature extraction using color snake, (2002), pp. 457–462. doi:10.1109/ISIE.2002.1026332
51. MH Mahoor, MA Mottaleb, AN Ansari, Improved active shape model for facial feature extraction in color images. *J. Multimedia.* **1**(4), 21–28 (2006)
52. SL Wang, WH Lau, AWC Liew, SH Leung, Robust lip region segmentation for lip images with complex background. *Pattern Recogn.* **40**, 3481–3491 (2007)
53. S Phimoltares, C Lursinsap, K Chamnongthai, Face detection and facial feature localization without considering the appearance of image context. *Image Vis. Comput.* **25**, 741–753 (2007)
54. E Bagherian, R Rahmat, in *Information Technology, 2008. ITSIM 2008. International Symposium on*, vol. 2. Facial feature extraction for face recognition: a review (Kuala Lumpur, Malaysia, 2008), pp. 1–9

55. D Xi, SW Lee, in *Proc. of Int. Conf. on Pattern Recognition*. Face detection and facial feature extraction using support vector machines, vol. 4, (2002), pp. 209–221
56. T Çelik, H Özkarmanlı, H Demirel, Facial feature extraction using complex dual-tree wavelet transform. *Comput. Vis. Image Understand.* **111**, 229–246 (2008)
57. RS Feris, J Gemmell, K Toyama, V Krüger, in *Proc. of Int. Conf. on Automatic Face and Gesture Recognition*. Hierarchical wavelet networks for facial feature localization (Washington, DC, USA, 2002), pp. 118–123
58. AA Salah, H Çınar, L Akarun, B Sankur, Robust facial landmarking for registration. *Annals Telecommun.* **62**(1–2), 83–108 (2006)
59. F Smeraldi, J Bigun, Retinal vision applied to facial features detection and face authentication. *Pattern Recogn. Lett.* **23**, 463–475 (2002)
60. EF Ersi, JS Zelek, in *Proc. of Int. Conf. on Image Analysis and Recognition*, vol. 3656. Rotation-invariant facial feature detection using gabor wavelet and entropy (Toronto, Canada, 2005), pp. 1040–1047
61. M Zobel, A Gebhard, D Paulus, J Denzler, H Niemann, in *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*. Robust facial feature localization by coupled features (Grenoble, France, 2000), pp. 2–7
62. GV Antonini, Popovici, JP Thiran, in *Proc. of Int. Conf. on Audio and Video-based Biometric Person Authentication*. Independent component analysis and support vector machine for face feature extraction (Guildford, UK, 2003), pp. 111–118
63. S Duffner, C Garcia, in *Compression et Représentation des Signaux Audiovisuels (CORESA)*. A Hierarchical Approach for Precise Facial Feature Detection (Rennes, France, 2005), pp. 29–34
64. HÇ Akakin, AA Salah, L Akarun, B Sankur, in *Proc. SPIE 6064*, Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning, 60641D (March 15, 2006). doi:10.1117/12.643099; [http://dx.doi.org/10.1117/12.643099]
65. R Brunelli, T Poggio, Template matching: matched spatial filters and beyond. Tech. Rep. Report No. 1549 MIT (1995)
66. GN Votsis, AI Drosopoulos, SD Kollias, A modular approach to facial feature segmentation on real sequences. *Signal Process. Image Commun.* **18**, 67–89 (2003)
67. K Sobottka, I Pitas, A novel method for automatic face segmentation, facial feature extraction and tracking. *Signal Process. Image Commun.* **12**, 263–281 (1998)
68. L Ding, AM Martinez, Features versus context: an approach for precise and detailed detection and delineation of faces and facial features. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(11), 2022–2038 (2010)
69. M Zhu, A Martinez, Subclass discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 1274–1286 (2006)
70. M Valstar, B Martinez, X Binefa, M Pantic, in *Proc. of Conf. on Computer Vision and Pattern Recognition*. Facial point detection using boosted regression and graph models (San Francisco, CA, USA, 2010), pp. 2729–2736
71. B Heisele, P Ho, J Wu, T Poggio, Face recognition: component-based versus global approaches. *Comput. Vis. Image Understand.* **91**, 6–21 (2003)
72. AL Yuille, D Cohen, P Hallinan, in *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*. Feature extraction from faces using deformable templates (San Diego, CA, USA, 1989), pp. 104–109
73. B Zhang, Q Ruan, in *Proc. of IEEE Int. Conf. on Signal Processing*, vol. 4. Facial feature extraction using improved deformable templates, (2006)
74. R Cesar, E Bengoetxea, I Bloch, in *Proc. of Int. Conf. on Pattern Recognition*, vol. 2. Inexact graph matching using stochastic optimization techniques for facial feature recognition (Quebec, 2002), pp. 465–468
75. Y Ryu, S Oh, Automatic extraction of eye and mouth fields from a face image using eigenfeatures and multilayer perceptrons. *Pattern Recogn.* **34**, 2459–2466 (2001)
76. TK Leung, MC Burl, P Perona, in *Proc. of Int. Conf. on Computer Vision*. Finding faces in cluttered scenes using random labeled graph matching (Boston, MA, USA, 1995), pp. 637–644
77. Z Zhu, Q Ji, in *Proc. of Int. Conf. on Pattern Recognition*, vol. 1. Robust pose invariant facial feature detection and tracking in real-time, (2006), pp. 1092–1095
78. GM Beumer, AMB Q Tao, RNJ Veldhuis, in *Proc. of Int. Conf. on Automatic Face and Gesture Recognition*. A landmark paper in face recognition (Southampton, UK, 2006)
79. T Cootes, G Edwards, C Taylor, in *Proc. of European Conf. on Computer Vision*, vol. 2. Active appearance models, (1998), pp. 484–498
80. D Cristinacce, T Cootes, in *Proc. of British Machine Vision Conference*, vol. 1. Facial feature detection using adaboost with shape constraints, (2003), pp. 231–240
81. D Cristinacce, T Cootes, in *Proc. of Int. British Machine Vision Conf.*, vol. 2. Boosted regression active shape models, (2007), pp. 880–889
82. D Cristinacce, T Cootes, I Scott, in *Proc. of Conf. on British Machine Vision*. A multi-stage approach to facial feature detection (Kingston University London, UK, 2004), pp. 277–286
83. D Cristinacce, T Cootes, Automatic feature localisation with constrained local models. *Pattern Recogn.* **41**, 3054–3067 (2008)
84. T Cootes, CJ Taylor, DH Cooper, J Grahama, Active shape models: their training and application. *Comput. Vis. Image Understand.* **91**, 38–59 (1995)
85. JM Saragih, S Lucey, JF Cohn, Deformable model fitting by regularized landmark mean-shift. *Int. J. Comput. Vis.* **91**, 200–215 (2011)
86. E Vezzetti, F Marcolin, 3D human face description: landmarks measures and geometrical features. *Image Vis. Comput.* **30**, 698–712 (2012)
87. B Gökberk, MO İrfanoğlu, L Akarun, 3D shape-based face representation and feature extraction for face recognition. *Image Vis. Comput.* **24**(8), 857–869 (2006)
88. X Lu, AK Jain, in *Proc. of Int. Conf. on Automatic Face and Gesture Recognition*. Automatic feature extraction for multiview 3D face recognition (Southampton, UK, 2006)
89. D Colby, G Stockman, A Jain, in *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*. Detection of anchor points for 3D face verification (San Diego, CA, USA, 2005), p. 118
90. H Dibeklioğlu, AA Salah, L Akarun, in *Proc. of IEEE Int. Conf. on Biometrics: Theory, Applications and Systems*. 3D facial landmarking under expression, pose and occlusion variations (Washington DC, 2008)
91. E Akagündüz, I Ulusoy, in *Proc. of Int. Conf. on Computer Vision*. 3D object representation using transform and scale invariant 3D features (Rio de Janeiro, Brazil, 2007), pp. 1–8
92. MP Segundo, L Silva, ORP Bellon, CC Queirolo, Automatic face segmentation and facial landmark detection in range image. *IEEE Trans. Systems Man Cybernetics—Part B: Cybernet.* **40**, 1319–1330 (2010)
93. P Nair, A Cavallaro, in *IEEE Trans. Multimedia*, vol. 11. 3D face detection, localization, landmark, and registration using a point distribution model, (2009), pp. 611–623
94. C Conde, LJ Rodríguez-Aragón, E Cabello, Automatic 3D face feature points extraction with spin images. *Image Anal. Recogn. LNCS*, Springer-Verlag. **4142**, 317–328 (2006)
95. AA Salah, L Akarun, 3D facial feature localization for registration. *Multimedia Content Representation, Classification and Security*. LNCS Springer Verlag. **4105**, 338–345 (2006)
96. FM Sukno, JL Waddington, PF Whelan, in *Computer Vision—ECCV 2012. Workshops and Demonstrations Lecture Notes in Computer Science*, vol. 7583. 3D facial landmark localization using combinatorial search and shape regression (Springer Berlin Heidelberg, 2012), pp. 32–41
97. X Cao, Y Wei, F Wen, J Sun, in *Proc. of Conf. on Computer Vision and Pattern Recognition*. Face alignment by explicit shape regression (Providence, RI, USA, 2012), pp. 2887–2894
98. T Kozakaya, T Shibata, M Yuasa, O Yamaguchi, in *Proc. of Int. Conf. on Automatic Face and Gesture Recognition*. Facial feature localization using weighted vector concentration approach (Amsterdam, The Netherlands, 2008), pp. 1–6
99. X Zhu, D Ramanan, in *Proc. of Conf. on Computer Vision and Pattern Recognition*. Face detection, pose estimation, and landmark localization in the wild (Providence, RI, USA, 2012), pp. 2879–2886
100. P Felzenszwalb, R Girshick, D McAllester, D Ramanan, Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9), 1627–1645 (2009)
101. A Torralba, KP Murphy, WT Freeman, Sharing visual features for multiclass and multiview object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(5), 854–869 (2007)
102. M Everingham, J Sivic, A Zisserman, in *Proc. of British Machine Vision Conf.* “Hello! My name is..Buffy”–automatic naming of characters in TV video (Edinburgh, 2006)

103. M Uricar, V Franc, V Hlavac, in *Proc. of Int. Conf. on Computer Vision Theory and Applications*, vol. 1. Detector of facial landmarks learned by the structured output SVM (Rome, Italy, 2012), pp. 547–556
104. M Dantone, J Gall, G Fanelli, LV Gool, in *Proc. of Conf. on Computer Vision and Pattern Recognition*. Real-time facial feature detection using conditional regression forests (Providence, RI, USA, 2012), pp. 2578–2585
105. B Efraty, C Huang, S Shah, I Kakadiaris, in *Proc. of Int. Joint Conf. on Biometrics*. Facial landmark detection in uncontrolled conditions (Washington, DC, 2011), pp. 1–8
106. PN Belhumeur, DW Jacobs, DJ Kriegman, N Kumar, in *Proc. of Conf. on Computer Vision and Pattern Recognition*. Localizing parts of faces using a consensus of exemplars (Providence, RI, USA, 2011), pp. 545–552
107. Y Tong, X Liu, FW Wheeler, PH Tub, Semi-supervised facial landmark annotation. *Comput. Vis. Image Understand.* **116**(8), 922–935 (2012)
108. V Rapp, T Senechal, K Bailly, L Prevost, in *Automatic Face Gesture Recognition and Workshops (FG 2011)*, 2011 *IEEE International Conference on*. Multiple kernel learning SVM and statistical validation for facial landmark detection (Santa Barbara, CA, USA, 2011), pp. 265–271
109. R Gross, I Matthews, J Cohn, T Kanade, S Baker, Multi-PIE. *Image Vis. Comput.* **28**(5), 807–813 (2010)
110. A Savran, N Alyüz, H Dibeklioğlu, O Çeliktutan, B Gökberk, B Sankur, L Akarun, in *Proc. of COST 2101 Workshop on Biometrics and Identity Management (BIOID)*, vol. 5372, ed. by B Schouten, J NielsChristian, A Drygajlo, and M Tistarelli. Bosphorus database for 3D face analysis (Springer Berlin Heidelberg, Lecture Notes in Computer Science Roskilde University, Denmark, 2008), pp. 47–56
111. A Martinez, R Benavente, The AR face database. Tech. Rep. 24 CVC Technical Report (1998)
112. K Messer, J Matas, J Kittler, J Luettin, G Maitre, in *Proc. of Audio- and Video-Based Person Authentication*. XM2VTS: the extended M2VTS database (Washington, DC, 1999)
113. T Kanade, JF Cohn, Y Tian, in *Proc. of Int. Conf. on Automatic Face and Gesture Recognition*. Comprehensive database for facial expression analysis (Grenoble, France, 2000), pp. 46–53
114. P Lucey, JF Cohn, T Kanade, J Saragih, Z Ambadar, I Matthews, in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010 *IEEE Computer Society Conference on*. The extended Cohn-Kanade dataset (CK+): a complete facial expression dataset for action unit and emotion-specified expression, (2010)
115. I Matthews, S Baker, Active appearance models revisited. *Int. J. Comput. Vis.* **60**, 135–164 (2004)
116. A Savran, B Sankur, T Bilge, 3D modality for automatic detection of facial action units. *Pattern Recogn.* **45**, 767–782 (2012)
117. P Phillips, H Wechsler, J Huang, P Rauss, The feret database and evaluation procedure for face-recognition algorithms. *Image Vis. Comput.* **16**(5), 295–306 (1998)
118. M Pantic, M Valstar, R Rademaker, L Maat, in *Proc. of Int. Conf. on Multimedia and Expo*. Web-based database for facial expression analysis (Amsterdam, The Netherlands, 2005), pp. 317–321
119. Lab, UCB, UHDB11 face database, 317–321 (2009). <http://cbl.uh.edu/URxD/datasets/>
120. O Jesorsky, KJ Kirchberg, RW Frischholz, in *Proc. of Conf. on Audio- and Video-Based Biometric Person Authentication*, Springer. Robust face detection using the Hausdorff distance (Hilton Rye Town, NY, USA, 2001), pp. 90–95
121. M Köstinger, P Wohlhart, PM Roth, H Bischof, in *Computer Vision Workshops (ICCV Workshops)*, 2011 *IEEE International Conference on*. Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization (Barcelona, Spain, 2011), pp. 2144–2151
122. G Huang, M Ramesh, T Berg, E Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep. Technical report, University of Massachusetts, Amherst (2007)
123. L Liang, R Xiao, F Wen, J Sun, in *Proc. of European Conf. on Computer Vision*, vol. 5303. Face alignment via component-based discriminative search (Marseille, France, 2008), pp. 72–85
124. PA Tresadern, H Bhaskar, SA Adeshina, CJ Taylor, TF Cootes, in *Proc. of British Machine Vision Conf.* Combining local and global shape models for deformable object matching (2009)
125. JM Saragih, S Lucey, JF Cohn, Deformable model fitting by regularized landmark mean-shift. *Int. J. Comput. Vis.* **91**, 200–215 (2011)
126. HA Rowley, S Baluja, T Kanade, Neural network-based face detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 23–38 (1998)
127. J Mairal, F Bach, J Ponce, G Sapiro, A Zisserman, in *Proc. of Conf. on Computer Vision and Pattern Recognition*. Discriminative learned dictionaries for local image analysis (Anchorage, Alaska, 2008), pp. 1–8
128. M Yang, L Zhang, X Feng, D Zhang, in *Proc. of Int. Conf. on Computer Vision*. Fisher discrimination dictionary for sparse representation (Barcelona, Spain, 2011), pp. 543–550
129. R Salakhutdinov, A Torralba, J Tenenbaum, in *Proc. of Conf. on Computer Vision and Pattern Recognition*. Learning to share visual appearance for multiclass object detection (Colorado Springs, USA, 2011), pp. 1481–1488
130. M Black, Y Yacoob, A Jepson, D Fleet, in *Proc. of Conf. on Computer Vision and Pattern Recognition*. Learning parameterized models of image motion (San Juan, Puerto Rico, 1997), pp. 561–567
131. H Wang, MM Ullah, A Kläser, I Laptev, C Schmid, in *Proc. of British Machine Vision Conf.* Evaluation of local spatio-temporal features for action recognition (London, UK, 2009), p. 127
132. D McDuff, R el Kaliouby, R Picard, Crowdsourcing facial responses to online videos. *IEEE Trans. Affective Comput.* **99**, 456–468 (2012)

doi:10.1186/1687-5281-2013-13

Cite this article as: Çeliktutan et al.: A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing* 2013 **2013**:13.

Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](http://springeropen.com)