A new framework for optimal facial landmark localization on light-field images

Light fields
The term “plenoptic” comes from the Latin words plenus ("full") + optic. The plenoptic function is the 5-dimensional function representing the intensity of the light observed from every point and direction in 3-dimensional space. Thanks to the plenoptic function it is thus possible to define the direction of every ray in the light field vector function.

Compared to classical 2D images, light fields capture the intensity values along each ray and not only the sum of intensities of rays reaching each image point:
- possibility of changing the angle of view;
- reconstructing the scene depth (depth map);
- and refocusing on the desired object in the scene.

Scan the QR code to read more and watch an animated GIF

Epipolar plane image (EPI)
The plenoptic camera captures a mosaic of micro images that after processing is translated into an array of images depicting the captured scene from different points of view (those of the micro lenses).

The EPI can be represented as a spatio-angular slice of the light field, cut through a horizontal (see the red line in the figure above) or vertical stack of light-field views.

Example of horizontal epipolar plane image (EPI): the pixel vectors corresponding to the horizontal red lines are stacked to form the EPI.

Observation
Following the inherent structure of light fields, the points laying on the same level line correspond to the projection, in the different angular (s) views, of the same 3D point in space.

If a given landmark detector performs likewise over all the views of the light-field face image, the detected landmarks should lay on the same level line in the EPI image.

Data
Data come from the IST-EURECOM Light Field Face Database, a face images database consisting of photos of 100 different persons, captured with a plenoptic camera in different face variations and illumination conditions. > http://iflid.eurecom.fr/

Method
The idea is thus to detect the level lines on which the detected face landmarks lie and use them to estimate the "true" level line. The coordinates of the points are then corrected accordingly.

Level lines are computed using the Structure Tensor.

Detected landmarks (blue points) and corresponding director vectors for the level lines computed from the structure tensor (green arrows).

**Def:** The structure tensor, also referred to as the second-moment matrix, is a matrix derived from the gradient of a function. It summarizes the predominant directions of the gradient in a specified neighborhood of a point, and the degree to which those directions are coherent.

Given a face landmark \( k \), let’s consider the black points in the figure below its position estimation over the views \( v_j \). For each black point, its level line is computed using structure tensor (black lines). The black points are then projected following the black lines to the central view (blue points, \( s_{v_j} \)). The coordinates of the point on the central view \( v_j \) are then corrected using weighted sum (green point):

\[
\hat{s}_{k,j} = \frac{\sum_{n \in N_{vi}} w_n s_{v_i} - \bar{c}_{v_j}}{\sum_{n \in N_{vi}} w_n}
\]

where the weight \( w_n \) is the number of black points at a distance less than 0.1 pixels form the corresponding level line.

Experiments
A set of light-field face images (50 subjects × 2 sessions × 4 poses variations = 400 images) has been selected from the IST-EURECOM Light Field Face Database.

For each light field, 15 horizontal and 15 vertical views have been extracted with regular angular sampling in the perspective range \([-0.5, 0.5]\) thanks to the LYTRO POWER TOOLS BETA. For a total of 12 000 images.

Faces on the central views have been manually annotated with 32 landmarks to constitute a set of ground-truth points.

Face landmarks are detected using DLIB, and then corrected with the proposed method. The original detected coordinates and the corrected ones are then evaluated using normalized root mean square error (NRMSE). DLIB > http://dlib.net/

Results
The NRMSE between the ground-truth coordinates \((x, y)\) and the estimated coordinates \((\hat{x}, \hat{y})\), is defined as:

\[\delta_k = \frac{1}{10D} \sum_{i=1}^{10D} \left[ \frac{|\delta_i^k|}{\delta_{max}} \right]_{\leq T_h}\]

where \(\delta_i^k\) indicates the Euclidean distance, \(k\) indicates the landmark index (e.g. eye corner, nose tip), \(i\) is the image angular coordinate (view), and 10D is the interocular distance.

The overall landmark detector performance in terms of percentage of detected landmarks, is computed by the following formula:

\[P \left( \frac{100}{10D} \sum_{i=1}^{10D} \sum_{k=1}^{16} [\delta_i^k < T_h] \right)\]

where \([\delta_i^k < T_h]\) is the indicator of value 1 if the distance is smaller than \(T_h\), 0 otherwise. \(I\) is the number of test images and \(K\) the number of landmarks per face image. And \(\delta\) is the NRMSE between the detected landmark and the ground truth.

Example of error threshold for \(T_h = 0.1\) (1/10 of 0D)
Scan the QR code to watch an animated GIF correction example >