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

Chiara Galdi [#1], Lara Younes [*2], Christine Guillemot [*3], Jean-Luc Dugelay [#4]

[#] *Digital Security Department, EURECOM*
450 Route des Chappes, CS 50193 - 06904 Biot Sophia Antipolis cedex, FRANCE
[1] chiara.galdi@eurecom.fr
[4] jean-luc.dugelay@eurecom.fr

[*] *Inria Rennes - Bretagne Atlantique*
Campus universitaire de Beaulieu - 35042 Rennes Cedex, FRANCE
[2] lara.younes@inria.fr
[3] christine.guillemot@inria.fr

***Abstract*—The paper explores how light fields captured by plenoptic cameras can increase the performance of face landmark detection. The idea is to exploit light fields geometrical constraints to correct the position of points detected by classical face landmark detectors. These geometric constraints are used to enforce landmark points angular coherency across the different views of the light field, and by doing so to correct the positions of the landmarks on all views. The corrected landmark points are compared with ground-truth manual annotations of a set of 400 images corresponding to the central views of 400 light fields of faces with different pose and expression.**

***Index Terms*—light fields, face landmark detection, structure tensor, LFFD, EPI**

I. Introduction

Imaging systems are rapidly evolving with the emergence of light fields capturing devices. As a consequence, existing image processing techniques need to be adjusted to suit the richer information provided. This paper explores how face landmark detection can be improved on light-field images by exploiting their particular structure and more precisely the so-called epipolar plane image (EPI).

Light fields have emerged as a representation of light rays emitted by a 3D scene and received by an observer at a particular point $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ in space, along different orientations. A variety of systems now exist for capturing real-world light fields which go from cameras arrays [1] to single cameras mounted on moving gantries and plenoptic cameras [2], [3]. Plenoptic cameras use an array of micro-lenses placed in front of the sensor to separate light rays which can then be recorded separately by different photosensors. Recent smart phones, equipped with either several cameras, with a single specialized sensor, or with a wafer-level-optics camera array [4], can also capture light fields. The recorded flow of rays, with varying angular coordinates t and s, gives a rich description of the scene which can benefit biometric applications.

In this paper, we explore how the scene geometry which is visible in epipolar plane images (EPI) of the light field can be used to improve the performance of landmark points detection in 2D views of the face. Landmarking points are detected in the multiple views of the light field captured by a plenoptic camera (in the experiments, light fields captured by a Lytro Illum camera have been used). The scene geometry extracted from the EPIs is used to enforce angular coherency of the detected points. The proposed approach exploits the fact that any 3D point of the scene is projected on 1D lines in the EPI. The slopes of these 1D lines correspond to the inter-view disparity, hence can be used to make sure that the detected landmark points are coherent across all the light field views. A method is investigated to estimate the 1D epipolar constraints, based on structure tensors.

The central views of a set of 400 light fields showing faces in different poses have been manually annotated with 32 landmarks which give the ground-truth landmark points. These landmarks correspond to points commonly extracted by state-of-the-art detectors and annotated databases, namely the AFLW (Annotated Facial Landmarks in the Wild) database [5]. The localization performance is measured in terms of root mean square distance, compared with the ground-truth positions, normalized by the Inter-Ocular Distance (IOD). Localization performances have been measured without and with correction using a state-of-the-art landmark detectors (DLIB[1]). Experimental results show a gain in localization precision.

In summary, the contributions of the paper are the following:

- We propose exploiting light field captures to improve the performance of landmark points detection.
- The proposed approach is independent of the algorithm used for detecting landmarks, and thus can potentially be adopted for correcting any face landmark estimation.
- The proposed framework and the set of annotated facial landmarks will be made available on request.

[1] http://dlib.net/

II. Related work

Face landmarking has been a very active field of research in the past decades and has known significant progress. A comprehensive overview of existing techniques can be found in [6]. Model based systems have proven to be efficient in facial landmark localization. Model-based landmark detectors [7], [8], learn the model of the face with a generative representation of its shape through a set of annotated landmarks and/or its appearance through the image texture. Their main goal is to find the model parameters that minimize the difference between a query face image and the face model.

Locally constrained methods have later been introduced. They rely on local models composed of discriminative features. The model is learned from discriminative feature descriptors computed on a local patch around the landmarks. Those local constraints of the features make the system less sensitive to illumination changes and possible self-occlusions.

In our experiments we use a locally constrained state-of-the-art landmark detector. DLIB is a modern C++ toolkit containing machine learning algorithms. It is used in both industry and academia in a wide range of domains including robotics, embedded devices, mobile phones, and large high performance computing environments. The face detector provided by this library, is made using the classic Histogram of Oriented Gradients (HOG) feature combined with a linear classifier, an image pyramid, and sliding window detection scheme. The pose estimator was created by using dlib's implementation of the paper [9], and was trained on the iBUG 300-W face landmark dataset[2] [10].

Despite significant advances in the past years, the experimental results reported here confirm that, under varying pose, expression, the detection performances degrade. To the best of our knowledge, this is the first work exploiting light field imaging characteristics for improving face landmark detection.

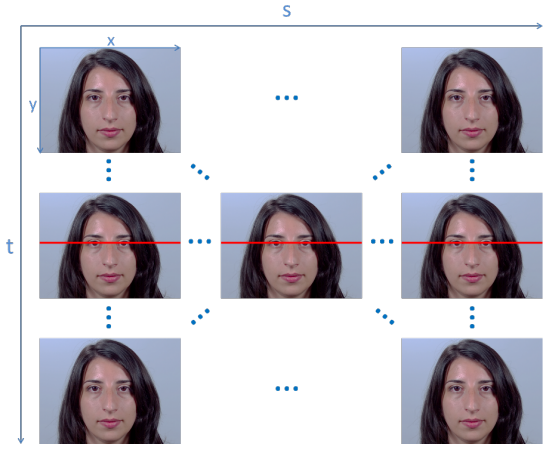
III. EPI and Structure Tensor

Epipolar plane image: The Epipolar plane image (EPIs) terminology has been introduced in [11]. It is derived from the epipolar constraint in stereo vision. The use of EPI has been explored for depth map construction and dense matching of images with narrow baseline.

The micro lenses being regularly spaced, the disparity of corresponding points (corresponding to the same 3D point) is the same between every pair of adjacent views. The disparity follows then a linear model over all the angular views allowing the exploration of the 3D geometry through the EPIs.

The EPI can be represented as a spatio-angular 2D slice of the 4D light field cut through a horizontal or vertical stack of light field views (see Fig. 1 (b)). They are obtained by fixing one of the spatial coordinates (e.g. y) and one of the angular coordinates (e.g. s). The EPI E_{y^*,t^*} shown in Fig. 1 (b) gives an observation of a 2D horizontal slice of the light field at a constant y^*-coordinate corresponding to the red line in the sub-aperture views of the 4D light field (Fig. 1 (a)). The slope

[2]https://ibug.doc.ic.ac.uk/resources/facial-point-annotations/



(a) A 4D representation of a light field.

(b) Epipolar plane image (EPI).

Fig. 1. Example of horizontal epipolar plane image (EPI).

of the level lines which can be observed along the s dimension of the horizontal EPIs can be computed locally for every (x, s) coordinate in E_{y^*,t^*}.

Following the inherent structure of the light field, 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.

In the following we suggest optimizing the landmark detection by relying on the detection results on all outer light field views (all views except for the central one).

Structure tensor: The structure tensor is defined as the second moment matrix and is computed from the EPI as:

$$J_\sigma(x,s) = \nabla E_{y^*,t^*}(x,s) \cdot \nabla E_{y^*,t^*}(x,s)^T * G_\sigma,$$

where G_σ is a Gaussian smoothing operator of variance σ^2. The orthogonal eigenvectors V_+ and V_- with respective eigenvalues λ_+ and λ_- (where $\lambda_+ > \lambda_-$) of $J_\sigma(x,s)$ give a robust computation of the local gradient orientations locally at (x, s) (see Fig. 2). In the EPI, we are interested in the eigenvector V_- with the smallest eigenvalue. It describes the director vector of the level lines $l(x^k_{v_s},\ V^k_{v_s\_})$ passing through (x, s).

Let F^k_v be the k^{th} facial landmark on the v^{th} view, where $k = 1, 2, \ldots, K$ and K is the number of detected DLIB landmarks, and $v = (s, t)$ where $s = 1, 2, \ldots, P$, $t = 1, 2, \ldots, Q$ and $V = P \times Q$ is the total number of extracted views (in our experiments we extract 15 vertical and 15 horizontal views). A landmark is defined by its coordinates $F^k_v = (x^k_v, y^k_v)$ in the corresponding view that in turn is defined by two coordinates $v = (s, t)$, indicating its position in the array of views. The idea is to correct landmark coordinates in the central view $v_c = (\lfloor \frac{P}{2} \rfloor, \lfloor \frac{Q}{2} \rfloor)$. The proposed method corrects the $x^k_{v_c}$ and

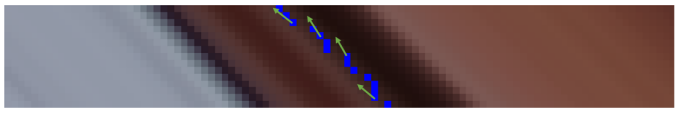


Fig. 2. Zoomed epipolar image: in blue, the x-coordinates of a detected landmark. In green, plot of the director vectors for the level lines computed from the structure tensor. Only some vectors have been drawn for a better visibility.

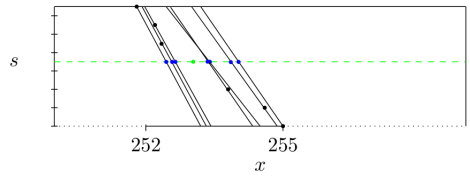


Fig. 3. Level lines (black) derived from the structure tensors and their intersection with the central view (dotted green line) for a given EPI. Black dots represent detected landmark points, while the blue dots represent the expected points.

$y^k_{v_c}$ coordinates separately using the horizontal and vertical slices of the light field array of images, respectively. We consider the horizontal and vertical slices (EPIs) of the light field composed of the views $v_s = (s, \lfloor \frac{Q}{2} \rfloor)$ and $v_t = (\lfloor \frac{P}{2} \rfloor, t)$ respectively, i.e. the ones including the central view v_c.

The x-coordinate of the landmark in the central view is computed as a weighted mean of the x-coordinates of the same landmark in the multiple angular views, with weights depending on the distance of the point from the estimated level line.

First, the weight $w_{v_s^k}$ for every coordinate in the angular views of the k^{th} landmark is determined as the number of x-coordinates in the multiple views that are within a distance of 0.1 pixel to its corresponding level line.

The landmark position in the central view is expected to lay at the intersection $\hat{x}^k_{s_{v_c}}$ of the level line of the landmark $F^k_{v_s}$ in the $E_{y*,\lfloor \frac{Q}{2} \rfloor}$ with the row corresponding to the central view at $\lfloor \frac{Q}{2} \rfloor$. The corrected location of the landmark in the central view is then computed as a weighted mean of those expected locations (see fig. 3) as follows:

$$x_{corr}{}^k_{v_c} = \frac{\sum_{s=0}^{P} w^k_{v_s} \cdot \hat{x}^k_{s_{v_c}}}{\sum_{s=0}^{P} w^k_{v_s}}.$$

The vertical slice of the light field is processed in a similar manner. In this case the $y_{corr}{}^k_{v_t}$ coordinate of the landmark is corrected.

IV. Experimental Results

As described in [6], a straightforward way to assess landmark detection performances is to compare the estimated points with manually annotated ground truth points. The localization performance can be expressed in terms of the normalized root mean square error (NRMSE). The normalization is typically done with respect to IOD: Inter-Ocular Distance, which is defined as the distance between the two eye centers. The normalization step is important for having performance measures independent of the image resolution or the actual face size in the picture. An error threshold, Th, is defined so that a point is said detected if its Euclidean distance from the ground truth is less that 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 [6]. In Fig. 4, three circles corresponding to $Th = 0.1,\ 0.05,\ 0.025$ are placed around a landmark. The performance metric is thus defined as the percentage of landmark points with a normalized Euclidean distance less than Th from the corresponding ground truth points. The normalized root mean square error between the ground truth coordinates (x, y) and the estimated coordinates $(\tilde{x}, \tilde{y})$, is defined as:

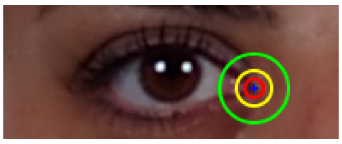


Fig. 4. Circles corresponding to different error thresholds: red corresponds to $Th = 0.025$; yellow to $Th = 0.05$; green to $Th = 0.1$.

$$\delta^k_v = \frac{d\{(x^k_v, y^k_v), (\tilde{x}^k_v, \tilde{y}^k_v)\}}{IOD} \tag{1}$$

where $d()$ indicates the Euclidean distance, k indicates the landmark index (e.g., eye corner, nose tip) and v is the image angular coordinate.

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

$$P = 100 \frac{\sum_{k=1}^{K} \sum_{i=1}^{I} [i : \delta^k_i < Th]}{K \times I} \tag{2}$$

where $[i : \delta^k_i < Th]$ is the indicator function of value 1 if the distance is smaller than Th, otherwise its value is 0. I denotes the number of test images and K the number of landmarks per face image.

In the presented experiments, we assessed the performances (P) using three thresholds: $Th = 0.1,\ 0.05,\ 0.025$.

A. Database

A set of face images (50 subjects $\times$ 2 sessions $\times$ 4 pose variations = 400 images) has been selected from the IST-EURECOM Light Field Face Database (LFFD)[12]. The images are captured with several facial variations. The first part of the database, captured at Instituto de Telecomunicações - Instituto Superior Técnico, Lisbon, Portugal can be accessed at http://www.img.lx.it.pt/LFFD/. The second part, captured at EURECOM, SophiaTech Campus, Nice, France can be accessed at http://lffd.eurecom.fr/. To assess the proposed framework for landmark correction from light fields, faces have been manually annotated with 32 landmarks to constitute a set of ground-truth points.

B. Experimental set up

This section summarizes information for reproducing the presented experiments. Landmark correction is performed on the light field horizontal and vertical slices/EPIs. The Lytro ILLUM camera has a very narrow baseline and computed disparities are below 1 pixel. For each light field, we extracted 15 horizontal and 15 vertical views thanks to the LYTRO POWER TOOLS BETA[3]. The views have been extracted with regular angular sampling in the perspective range $[-0.5, 0.5]$. The tool allows sampling in the range $[-1.0, 1.0]$ but large perspective changes can result in artifacts. The use of the above-mentioned settings is recommended for a correct computation of local gradient orientation.

Four face variations have been selected for testing, from less to more challenging: neutral, open mouth, look upward, and half profile. For each face variation, 100 light fields have been selected from the LFFD. Considering that for each light field image 30 views have been extracted, in our experiments, a total of 12.000 images have been used. To assess performances, 400 central views (4 variations $\times$ 100 light fields) have been manually annotated.

C. Performance gain

The results obtained are summarized in table I. An overall improvement is observed over the different poses/expressions for different thresholds. The threshold typically used for landmark localization performance assessment is $Th = 0.1$ and in this case the corrected localization performs better for all face variations. We tested smaller thresholds as well to analyse the gain in precision at a finer level. The performance gain is remarkably interesting for the half profile (up to 2.36% and of 3.99% overall). This face pose is extremely challenging for landmark detectors.

The proposed method is particularly beneficial for improving landmark localization on challenging face variations (see the overall gain in table I). That is desirable since existing detectors perform poorly on strong face variations - as demonstrated by the results for the original localization.

Regarding processing speed, the most time-consuming operation is the computation of the eigenvectors of the structure tensor over the entire EPI. However, this computation can be restricted to the areas where landmark points have been detected, and it is easily parallelizable, making the processing in interactive time feasible.

TABLE I
PERFORMANCE ASSESSMENT ON DIFFERENT FACE VARIATIONS

	P (%) Th = 0.1	P (%) Th = 0.05	P (%) Th = 0.025	Overall gain (%)
Neutral Frontal Face				
Original	97.81	77.30	44.06	- 0.25
Corrected	98.11	77.76	43.05	
Action Mouth Open				
Original	95.70	67.76	28.31	1.64
Corrected	96.37	68.24	28.80	
Pose Up Looking				
Original	91.80	66.46	31.46	1.96
Corrected	92.66	66.34	32.68	
Pose Half-profile Left				
Original	77.68	43.56	18.94	3.99
Corrected	79.13	45.92	19.12	

V. CONCLUSION

In this paper, we have demonstrated how the scene geometry, which is visible in epipolar plane images (EPI) of the light field, can be used to improve the performance of landmark detection. The method is tested on a large set of images and on several face variations. The results show that improvements up to 2% are obtained after landmark correction.

Apart from the accuracy improvement, the method also addresses the case in which the face is misdetected in one or more of the light-field views. The landmarks of the misdetected face can be estimated from the surrounding views exploiting the scene geometry which is visible in epipolar plane images (EPI).

The proposed approach is fully reproducible and suitable for correcting landmarks detected with any approach. As an additional contribution, the manual annotation of the LFFD database will be provided on request.

[3]https://www.lytro.com/imaging/power-tools

REFERENCES

[1] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," in *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3. ACM, 2005, pp. 765–776.

[2] R. Ng, "Digital Light Field Photography," Ph.D. dissertation, Stanford University, Stanford, CA, USA, 2006.

[3] T. Georgiev, G. Chunev, and A. Lumsdaine, "Superresolution with the focused plenoptic camera," in *Computational Imaging*, 2011.

[4] C. Huang, H.-H. Chin, Y. Wang, and L. Chen, "Fast realistic refocusing for sparse light fields," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*, 2015, pp. 1176–1180.

[5] M. Kstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Nov 2011, pp. 2144–2151.

[6] O. Çeliktutan, S. Ulukaya, and B. Sankur, "A comparative study of face landmarking techniques," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 13, Mar 2013.

[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active Shape Models-Their Training and Application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Computer Vision ECCV98*, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, Jun. 1998, pp. 484–498.

[9] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1867–1874.

[10] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *2013 IEEE International Conference on Computer Vision Workshops*, Dec 2013, pp. 397–403.

[11] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *International Journal of Computer Vision*, vol. 1, no. 1, pp. 7–55, Mar. 1987.

[12] A. Sepas-Moghaddam, V. Chiesa, P. L. Correia, F. Pereira, and J. L. Dugelay, "The ist-eurecom light field face database," in *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, April 2017, pp. 1–6.