INTEGRATION OF PRECISE IRIS LOCALIZATION INTO ACTIVE APPEARANCE MODELS FOR AUTOMATIC INITIALIZATION AND ROBUST DEFORMABLE FACE TRACKING

Sebastian Vater, Ralph Ivancevic and Fernando Puente León

Institute of Industrial Information Technology (IIIT), Karlsruhe Institute of Technology (KIT),
76187 Karlsruhe, Germany

ABSTRACT

Face tracking and facial landmarking plays an important role in human-machine interaction. Active-Appearance Models (AAM) represent a well-established method to mathematically describe face characteristics such as facial expression and attentiveness. However, the applicability of AAM suffers from their high sensitivity to initialization and low robustness in scenarios of strong head movements. We show how a robust, automatic initialization of AAM can be performed by integration of precise iris localization into the AAM fitting algorithm. The shape and global transform parameters are extracted from the re-initialized model to provide well-suited start parameters for the subsequent fitting. The method is validated on the *gi4e* database showing an improvement to naive initialization and re-initialization in terms of robustness and accuracy.

Index Terms— Computer vision, active appearance models, facial landmarking, eye localization, face tracking

1. INTRODUCTION

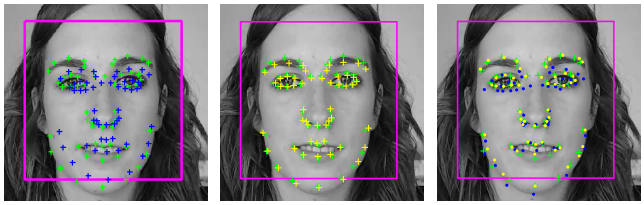
Face tracking and facial landmarking under uncontrolled conditions utilizing deformable models remains a challenging task and is widely discussed in recent machine vision research [1, 2]. Deformable trackers can be divided into discriminative models, that often learn a regression function to fit an object's appearance to the model [3–8], and generative methods, that model the object directly by minimizing an objective function with respect to model parameters [9–12]. Here, we follow the generative approach of the well-known Active Appearance Models (AAM) [9]. AAM have been utilized to successfully perform face tracking and facial landmarking [13, 14], were extended to multi-view models [15] and to perform reconstruction of the 3D head pose [16]. Constrained local models (CLM) [3, 5] have proven to generalize better on unseen images by correlation-based matching of local feature templates.

A minimization of the objective function is performed using the Gauss-Newton algorithm in [17] resulting in the fast inverse compositional implementation. The method in [17] is extended to a combined forward additive and inverse compositional method in [18]. In [12], Active Orientation Models (AOM) are presented that utilize a robust kernel principal component analysis (PCA) employing an orientation-based appearance model, while in [19] AAM are combined with Histogram of Oriented Gradients (HOG) features to obtain a better generalizing model. In [19], a drop in performance is reported for differing training and test data sets as well as for inaccurate initializations by mere face detection. A discriminative local model utilizing a Bayesian approach employing local support vector machine (SVM) detectors and a trained global model is presented in [4].

However, AAM suffer from stability problems under large head rotations and are especially sensitive to initialization inaccuracies [2, 12, 14]. At present, AAM and other facial landmarking algorithms have in common that they are initialized manually, with a-priori known ground truth or by re-initialization based on the previous frame [14, 20], with a reported increase of accuracy by about 20 % in [2].

To cope with this problem, rather than manually selecting fiducial points or following a naive approach that relies on empirically found anthropometric values to translate and scale the model to best fit a bounding box, we utilize a precise localization of the iris centers [21] and incorporate this information into the AAM fitting process. The presented method is capable of automatic, robust initialization of AAM, allowing for face tracking without any additional user action, making it applicable to any human-computer system.

We show how, after an affine-based transformation that incorporates the 2D iris information, the shape and global transform parameters of the AAM can be reconstructed from the model estimate. By decomposing the iris-corrected current model estimate, we retain the actual shape parameters utilizing our implementation of the inverse compositional algorithm [17]. We investigate how a frame-wise re-initialization affects the tracking results if the shape parameters are recovered after correcting the model estimate of the preceding frame by the iris information. Experiments with automatic model instantiation are conducted on the *gi4e* database [20]. Furthermore, we show how face tracking benefits by re-initializing the AAM frame-wise with the proposed method while affecting the computational speed to a minimal extent.



(a) (b) (c)

Fig. 1: First frame: *Naive* (a) (blue crosses), *Iris Recovered* (b) (yellow crosses) and the tracking results (c) (blue and yellow dots). The ground truth is shown in green and the bounding box of the detector in magenta. It should be noted, that the *Iris Recovered* initialization does not differ from *Naive Iris* since no shape parameters from a previous frame exist. Especially the regions around the eyes and the chin show a better tracking result with the proposed initialization.

(a) (b) (c)

Fig. 2: Amidst a sequence: *Naive* (a), *Naive Iris* (b) (both re-init at this frame due to poor tracking results in the previous frame) and *Iris Recovered Frame-wise* (c) depicted in blue, cyan and red. It is clearly visible that the initialization procedure benefits from exploiting the warp composition as described.

2. MODEL FORMULATION

Based on the original presentation of AAM that consists of statistical models of appearance and texture, we define the shape of an object in the object frame by $\mathbf{s} = [u_1^0, v_1^0, \ldots, u_\nu^0, v_\nu^0]^{\mathrm{T}}$, where ν is the number of model points, $\mathbf{u} = [u, v]^{\mathrm{T}}$ is a pixel coordinate and $g(\mathbf{u})$ a gray-valued image. The shape $\mathbf{s}(\mathbf{p})$ is parametrized by its shape parameters p_l via $\mathbf{s}(\mathbf{p}) = \mathbf{s}_0 + \sum_{l=1}^{n} p_l \mathbf{s}_l$, where $\mathbf{s}_0$ is the zero-mean so-called mean shape of the object shape model and n the number of shape vectors $\mathbf{s}_l$. The shape vectors are extracted utilizing PCA from a Procrustes-aligned training set, where the first n eigenvectors are taken into account for the model, with $\mathbf{s}_0$ being the zeroth eigenvector. Since $\mathbf{s}_1 \ldots \mathbf{s}_\nu$ only comprise the object's intra class (e. g. faces) shape variation, a global affine transformation $\mathbf{s}(\mathbf{q}) = \mathbf{s}_0 + \sum_{j=1}^{4} q_j \mathbf{s}_*$, with parameters q_j is applied, $\mathbf{s}_*$ being the scale, 2D translation and rotation vectors. We denote the shape transformation as $M(\mathbf{u}, \mathbf{p})$ and the global transformation as $N(\mathbf{u}, \mathbf{q})$, yielding the coordinates $\mathbf{u} = N(M(\mathbf{u}, \mathbf{p}), \mathbf{q})$ in the image frame, which we call the warp composition for the modeled object, following the notation in [17]. The texture variation is also modeled linearly by $\mathbf{A}(\mathbf{u}, \boldsymbol{\lambda}) = \mathbf{A}_0(\mathbf{u}) + \sum_{k=1}^{m} \lambda_k \mathbf{A}_k(\mathbf{u}) \;\; \forall\, \mathbf{u} \in \mathbf{s}_0$, where λ_k are the texture parameters. The eigenvectors $\mathbf{A}_k(\mathbf{u})$ are again extracted via PCA from the training images by transforming the actual shapes $\mathbf{s}_{\mathrm{train}}$ into $\mathbf{s}_0$. Note, that the ground truth landmarks $\mathbf{s}_{\mathrm{train}}$ must be known. The objective function if formulated following the project-out inverse compositional algorithm [17] comprising the error image over all pixel locations within the mean shape:

$$\arg\min_{\mathbf{p},\mathbf{q}} \sum_{\mathbf{u} \in \mathbf{s}_0} \left(\mathbf{A}_0(\mathbf{u}) - g\left(N(M(\mathbf{u}, \mathbf{p}), \mathbf{q})\right)\right) . \qquad (1)$$

After linearization , the objective function is iteratively minimized to solve for incremental $[\Delta p_l, \Delta q_j]^{\mathrm{T}}$, see [17] for details.

2.1. Fitting

Utilizing the inverse compositional algorithm to compute the incremental parameter changes $[\Delta p_l, \Delta q_j]^{\mathrm{T}}$, in each iteration of the fitting algorithm the actual warp $N(M(\mathbf{u}, \mathbf{p}), \mathbf{q})$ is applied to the input image to compute the error image. Thus, the warp composition $N(M(\mathbf{u}, \mathbf{p}), \mathbf{q})$ must be updated iteratively by the computed incremental parameter changes $[\Delta p_l, \Delta q_j]^{\mathrm{T}}$ by composing it inversely with the actual estimate

$$N(M(\mathbf{u}, -\Delta\mathbf{p}), -\Delta\mathbf{q}) \cdot N(M(\mathbf{u}, \mathbf{p}), \mathbf{q}) . \qquad (2)$$

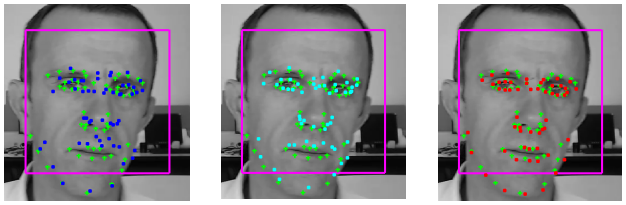
The convolutions above reduce to a multiplication for affine transformations. At the beginning of each iteration, the parameters $\mathbf{q}, \mathbf{p}$ must then be recovered from the composition.

2.2. Precise Iris Center Localization

Our implementation of iris localization follows the method presented in [21] that has shown to provide precise results while being inexpensive to compute. The approach utilizes isophotes, which can be understood as contours of constant intensity $g\left(\boldsymbol{\gamma}(\mathbf{u})\right) = \mathrm{const.}$ along a curve $\boldsymbol{\gamma}(\mathbf{u})$. Utilizing the local radius $r(\mathbf{u})$, a displacement vector is computed for each pixel by multiplying it with the local gradient: $\mathbf{d}(\mathbf{u}) = r(\mathbf{u}) \cdot \left(g_u(\mathbf{u}), g_y(\mathbf{u})\right)^{\mathrm{T}}$. The iris center is then estimated by summing over all displacement vectors while weighting the start- and endpoints with the curvedness and a gray value-based measure, respectively. The method yields estimates for the two irises $\mathbf{u}^{\mathrm{i}} = (\mathbf{u}^{\mathrm{i}}_{\mathrm{left}}, \mathbf{u}^{\mathrm{i}}_{\mathrm{right}})$ that can be incorporated into the AAM fitting algorithm to provide a suitable initial position of the AAM model.

3. INTEGRATING IRIS INFORMATION INTO AAM

Rather than relying on empirically found values for (re-)initialization of the mean shape $\mathbf{s}_0$, one could naively scale, translate and rotate the shape $\mathbf{s}(\mathbf{p}, \mathbf{q})$ with the iris centers by an affine transformation $L(\mathbf{u}^{\mathrm{i}})$ to obtain a new shape $\mathbf{s}(\mathbf{p}, \mathbf{q}, \mathbf{u}^{\mathrm{i}}) = L(\mathbf{u}^{\mathrm{i}}) \cdot \mathbf{s}(\mathbf{p}, \mathbf{q})$. This can be a proper approach



(a) (b) (c)

Fig. 3: Tracking results based on the initializations shown in Fig. 2.

Table I: MAE (Euclidean distance) of tracking results.

User	*Naive*	*N. Iris*	*Iris Rec.*	*Iris Rec. F.*	Ariz [20]
1	13.01	2.84	2.85	1.97	-
2	43.68	5.09	5.16	4.04	-
3	22.50	3.71	3.71	4.42	-
4	13.33	8.57	7.94	4.17	-
5	30.25	10.71	10.23	2.66	-
6	11.75	10.02	7.70	4.34	-
7	16.87	17.66	16.56	18.34	-
8	20.20	4.58	4.26	3.68	-
9	7.43	4.42	4.40	2.81	-
10	10.88	5.45	5.87	4.77	-
Avg.	17,87	6.06	5.61	3.83	6.70

when a new initialization is necessary (e. g. the first frame) and no information about $\mathbf{p}$ is known or $\mathbf{p}$ equals zero. In this case, an adjustment of the shape utilizing $\mathbf{u}^{\mathrm{i}}$ can be understood as a warp with $\mathbf{q} \neq 0$ and $\mathbf{p} = 0$. However, while tracking a face over many frames, the shape parameters $\mathbf{p}$ contain the present intra class variation of the actual face and should be used for further tracking. For this reason, to properly compose the warp at the beginning of each iteration, the information that is contained in $\mathbf{s}(\mathbf{p})$ must be recovered, since it differs from $\mathbf{s}_0$ in contrast to the case of a new initialization with an affine transformed mean shape.

4. RECOVERY OF SHAPE AND GLOBAL PARAMETERS

To incorporate $L(\mathbf{u}^{\mathrm{i}})$ into the warp composition, we extract the global transform parameters by computing the inner product with the global transform vectors

$$\mathbf{q}^{\mathrm{i}} = \langle \mathbf{s}_*, \mathbf{s}_0 - \mathbf{s}(\mathbf{p}, \mathbf{q}, \mathbf{u}^{\mathrm{i}}) \rangle \tag{3}$$

to obtain the zero-mean shape

$$\mathbf{s}(\mathbf{p}, \mathbf{q}^{\mathrm{i}}) = N(\mathbf{q}^{\mathrm{i}}) \cdot \mathbf{s}(\mathbf{p}, \mathbf{q}, \mathbf{u}^{\mathrm{i}}) \,. \tag{4}$$

We then calculate the inner product of the result with the shape vectors to recover the shape parameters that correspond to the correction of the actual shape estimate $\mathbf{p}, \mathbf{q}$ with the iris center estimation:

$$\mathbf{p}^{\mathrm{i}} = \langle \mathbf{s}_l, \mathbf{s}_0 - \mathbf{s}(\mathbf{p}.\mathbf{q}^{\mathrm{i}}) \rangle. \tag{5}$$

By employing $N(M(\mathbf{u}, \mathbf{p}^{\mathrm{i}}), \mathbf{q}^{\mathrm{i}})$ when updating the warp after estimating the next parameter update, we retain the composition of the shape and global parameters according to the last iteration. Note, that two problems might occur when neglecting the recovery of the composition of the warp (see results section):

(1) The transformation $\mathbf{s}(\mathbf{p}, \mathbf{q}) \cdot L(\mathbf{u}^{\mathrm{i}})$ not necessarily equals $N(M(\mathbf{u}, \mathbf{p}^{\mathrm{i}}), \mathbf{q}^{\mathrm{i}})$. While the latter represents a true model instance, the former one is any affine transformation that might not be covered by the model space.

(2) When employing a new model estimate for the initialization based on $L(\mathbf{u}^{\mathrm{i}})$ with unknown parameters $\mathbf{p}, \mathbf{q}$, the warp update in (2) must be applied with $\mathbf{p} = 0$ as L only consists of a global transformation. Therefore, the actual shape parameters are not considered (which is done in case of re-initialization with $\mathbf{s}(\mathbf{p}, \mathbf{q})$ with the previous frame estimate) in the warp update.

5. RESULTS

To validate our method and to show its applicability, we conducted experiments on the *gi4e* database that consists of 10 subjects and 12 videos each, showing different head movements [20]. Since we are particularly interested in evaluating the presented method, we validate our approach by applying person specific AAM to account for the AAMs limited ability to generalize to unknown data [10]. We trained a model by manually selecting between 12 and 21 images of each individual while retaining 90 % of the variance.

5.1. Naive initialization *(Naive)*

The first evaluation follows the naive approach and uses the bounding box of the OpenCV face detector [22] and empirically found values to initialize the model. We found that scaling the mean shape by 0.7 of the box size yields satisfactory results which we validated by visual inspection.

The naive initialization and the ground truth of the first frame of a video are depicted in Fig. 1(a). Fig. 2(a) shows the initialization for a frame amidst a video sequence. Initialization is performed in the first frame of each video and re-initialized if at least 7 of the 54 landmark points lie outside the detector's bounding box. For all other frames, the previous estimate of the model is used at the beginning of the first iteration of the method.

5.2. Automatic Initialization *(Naive Iris)*

In a next step, we investigate the influence of an automatic initialization utilizing the iris centers, but without recovering the shape and global transform parameters as described in Section 4. In this case, we apply the affine transformation $L(\mathbf{u}^{\mathrm{i}})$ to the mean shape by computing $\mathbf{s}_0 \cdot L(\mathbf{u}^{\mathrm{i}}) = \mathbf{s}(0, \mathbf{q})$ at initialization. Since this implies minimizing (1) starting with $\mathbf{p} = 0$, it may occur that the algorithm tries to find $\Delta\mathbf{p}$ that actually drives the shape against $\mathbf{q}$, stimulating the algorithm

(a) (b) (c) (d) (e)

Fig. 4: Comparison of the different methods, user 6: Mean appearance $\mathbf{A}_0$ (Fig. 4(a)). Results on video 8: 4(b): *Naive* (blue), 4(c): *Naive Iris* (cyan), 4(d): *Iris Recovered* (yellow) and 4(e): *Iris Recovered Frame-wise* (red). The ground truth is depicted in green.

to diverge, since $\Delta\mathbf{p}$ might fail its purpose to explain shape changes rather than global changes. Initialization following *Naive Iris* is shown in cyan in Fig. 2(b).

5.3. Automatic Initialization with Recovery *(Iris Recovered)*

To incorporate the model change due to the transformation utilizing the iris positions into the algorithm, we apply L as before but follow the steps in (3), (4) and (5) thereafter. This way, the actual shape can be retained ($\mathbf{p} = 0$ before) while the model manipulations are expressed in both $\mathbf{p}$ and $\mathbf{q}$. This method is especially interesting when the face is non frontal and shows a facial expression different from the mean shape.

5.4. Automatic Initialization with Frame-wise Recovery *(Iris Recovered Frame-wise)*

We want to investigate a possible benefit from further refining the model estimate by applying (3), (4) and (5) frame-wise. Initialization and tracking results with *Iris Recovered Frame-wise* are shown in Figs. 2(c) and 3(c), respectively. To exploit the capability of recovering $\mathbf{p}$ and $\mathbf{q}$, Fig. 4 shows a rotated face where the user is smiling. The results for the approaches discussed in Sections 5.1, 5.2, 5.3 and 5.4 are depicted in Figs. 4(b), 4(c), 4(d) and 4(e).

5.5. Discussion

Table I summarizes the results on the *gi4e* database. By initializing the AAM utilizing the iris information the accuracy is increased by more than 10 pixels in average for each method compared to a naive initialization. This can be explained by an appropriate adaption of the mean shape to the current image utilizing person-based as well as situation-based anthropometric values gained from the iris positions rather than adopting a general empirical value for each user. The method allows for generic application of facial landmarking by enabling an automatic AAM initialization.

While the *Naive Iris* method already improves tracking dramatically, further improvements can be observed by recovering the shape and global parameters and updating the estimate, demonstrating the ability of the presented method, furthermore showing an improvement of more than $30\,\%$ between *Iris Rec.* and *Iris Rec. Frame-wise*. This can be explained because the former method affects the results only in the case of (re-)initialization, while the latter refines the preceding estimate continuously. The poor results for user 7 originate from a failing tracking in video 7, where the model malfunctions. By excluding video 7 from the average, an MAE of $5{,}46$ is achieved for user 7. Utilizing cross-validation and a different implementation of the AAM fitting algorithm, the authors in [20] achieve an average MAE of 6.70, while initializing with GT data and using the previous estimate in subsequent frames.

The capability of the presented method can be interpreted by the fact that solving the objective function utilizing the decomposition of the adapted initialization into $\mathbf{p}$ and $\mathbf{q}$ enhances the convergence into the appropriate minimum. This is demonstrated in Figs. 4(c) and 4(d) , where the *Naive Iris* and *Iris Rec.* approaches fail, while the frame-wise method yields accurate tracking results.

6. CONCLUSION

In this paper we presented a method that allows for automatic initialization and precise re-initialization of AAM. The approach can be utilized in human-machine systems without requiring further user interaction making deformable face tracking independent and autarkically applicable. By incorporating precise iris center information into the warp composition of the AAM algorithm, an increased accuracy for problem-suited models could be shown.

To further illustrate its potential, another investigation should be performed on a dataset that represents a wide scope of facial expressions exploiting the strength of the proposed method lying in the decomposition of shape and global parameters. We believe that the method can be applied to other parameter-based facial landmarking algorithms to increase their performance and to allow for automatic initialization and therefore for better usability.

7. REFERENCES

[1] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Int. Conf. Computer Vision Workshop*. 2015, pp. 1003–1011, IEEE.

[2] G.G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou, "A comprehensive performance evaluation of deformable face tracking "in-the-wild"," in *arXiv preprint arXiv:1603.06015*, 2016.

[3] D. Cristinacce and T. F. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, 2008.

[4] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2930–2940, 2013.

[5] G. Rajamanoharan and T. F. Cootes, "Multi-view constrained local models for large head angle facial tracking," in *Proc. IEEE Int. Conf. Computer Vision Workshops*, 2015, pp. 18–25.

[6] Y. Wu and Q. Ji, "Shape augmented regression method for face alignment," in *Proc. IEEE Int. Conf. Computer Vision Workshops*, 2015, pp. 26–32.

[7] S. Xiao, S. Yan, and A. A. Kassim, "Facial landmark detection via progressive initialization," in *Proc. IEEE Int. Conf. Computer Vision Workshops*, 2015, pp. 33–40.

[8] J. Yang, J. Deng, K. Zhang, and Q. Liu, "Facial shape tracking via spatio-temporal cascade shape regression," in *Proc. IEEE Int. Conf. Computer Vision Workshops*, 2015, pp. 41–49.

[9] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," in *Eur. Conf. Computer Vision*. Springer, 1998, pp. 484–498.

[10] R. Gross, I. Matthews, and S. Baker, "Generic vs. person specific active appearance models," *Image and Vision Computing*, vol. 23, no. 12, pp. 1080–1093, 2005.

[11] X. Liu, "Generic face alignment using boosted appearance model," in *Conf. Computer Vision and Pattern Recognition*. 2007, pp. 1–8, IEEE.

[12] G. Tzimiropoulos, J. Alabort-i Medina, S. P. Zafeiriou, and M. Pantic, "Active orientation models for face alignment in-the-wild," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2024–2034, 2014.

[13] D. Cristinacce and T.F. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, 2008.

[14] G. Tzimiropoulos and M. Pantic, "Gauss-newton deformable part models for face alignment in-the-wild," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 1851–185.

[15] T.F. Cootes, G.V. Wheeler, K.N. Walker, and C.J. Taylor, "Coupled-view active appearance models," in *Brit. Machine Vision Conf.*, 2000, pp. 52–61.

[16] J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-time combined 2d+ 3d active appearance models," in *Conf. Computer Vision and Pattern Recognition*, 2004, pp. 535–542.

[17] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[18] J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "Fast and exact bi-directional fitting of active appearance models," in *Int. Conf. on Image Processing*. 2015, pp. 1135–1139, IEEE.

[19] E. Antonakos, J. Alabort-i Medina, G. Tzimiropoulos, and S. Zafeiriou, "HOG active appearance models," in *Int. Conf. Image Processing*. 2014, pp. 224–228, IEEE.

[20] M. Ariz, J.J. Bengoechea, A. Villanueva, and R. Cabeza, "A novel 2d/3d database with automatic face annotation for head tracking and pose estimation," *Computer Vision and Image Understanding*, vol. 148, pp. 201–210, 2016.

[21] S. Vater and F. Puente León, "Combining isophote and cascade classifier information for precise iris localization," in *IEEE Int. Conf. Image Processing*, 2014, pp. 1851–185.

[22] "Open source computer vision library," http://sourceforge.net/projects/opencvlibrary/.