2018

Looking at faces in the wild

Eugene Borovikov

Szilárd Vajda

Michael Bonifant

Michael Gill

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 123 (2018) 104–109

Procedia Computer Science

www.elsevier.com/locate/procedia

8th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2017

Looking at faces in the wild

Eugene Borovikov[1*], Szilárd Vajda[2†], Michael Bonifant[1‡], and Michael Gill[1]

[1]*National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*

[2]*Department of Computer Science, Central Washington University, WA, USA*

{Eugene.Borovikov, Michael.Bonifant, MGill}@NIH.gov, Szilard.Vajda@CWU.edu

Abstract

Recent advances in the face detection (FD) and recognition (FR) technology may give an impression that the problem of face matching is essentially solved, e.g. via deep learning models using thousands of samples per face for training and validation on the available benchmark data-sets. Human vision system seems to handle face localization and matching problem differently from the modern FR systems, since humans detect faces instantly even in most cluttered environments, and often require a single view of a face to reliably distinguish it from all others. This prompted us to take a biologically inspired look at building a cognitive architecture that uses artificial neural nets at the face detection stage and adapts a *single image per person* (SIPP) approach for face image matching.

Peer-review under responsibility of the scientific committee of the 8th Annual International Conference on Biologically Inspired Cognitive Architectures

Keywords: face detection, face matching, artificial neural network, single image per person

1 Introduction

Faces play a critical role in social interactions presenting a very convenient and non-intrusive way for visual identification and non-verbal communication. Although recent research on macaques indicates that facial identity may be encoded via a simple neural code that relies on the ability of neurons to distinguish facial features along specific axes in face space [1], we still do not understand how humans detect and read faces with little visual sampling per individual, generalizing their recognition ability to a vast variety of lighting, poses and expressions.

Modern face recognition (FR) systems have become quite advanced in recent years, showing near-human abilities to recognize faces [2]–[4] on very challenging face datasets [5]–[7]. Nearly all of them rely on deep neural nets (DNN), whose we recently observed due to the availability of affordable graphics processing units (GPU) allowing to train DNNs in hours rather than days.

DNNs originally have been inspired by biological perceptual systems [8] and have been shown to solve complex pattern recognition problems [9], but they appear to learn statistical patterns very

[*] Mastermind behind the approach and system integration

[†] Crucial methodology research and development

[‡] Essential implementation components and evaluation

Peer-review under responsibility of the scientific committee of the 8th Annual International Conference on Biologically Inspired Cognitive Architectures

10.1016/j.procs.2018.01.017

differently from humans, as primates typically require just a few visual samples of an object, to start recognizing it from various view-points, while their artificial counterparts require thousands of samples per object to start approaching human-level recognition accuracy.

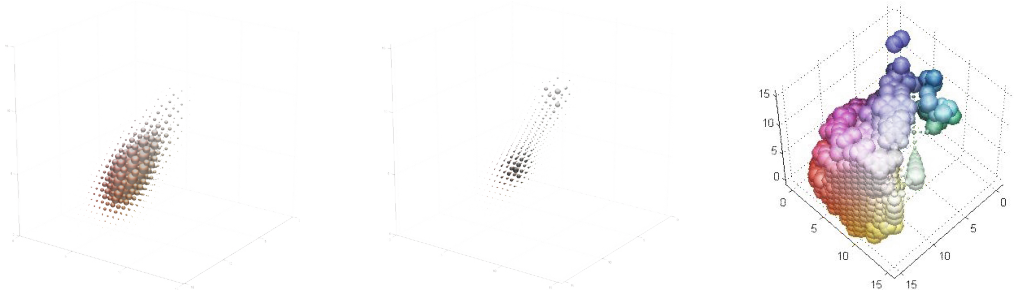
That prompted us to research and develop (R&D) a light weight (yet accurate) face detector and a *single image per person* (SIPP) face matcher, which is less complex than modern DNN systems, yet it is able to (a) use a single visual sample per subject, (b) be comparably accurate on unconstrained images, (c) adapt to the test visuals, and (d) run in near real-time requiring minimal computing power.

Our method cannot claim near-human level detection or recognition accuracy, but it does use several biologically inspired elements and it is utilized in a real-world face image retrieval system [10]. As biological systems inherit and then build up their perceptual abilities from the sensory experience, we proceed by R&D of a data-driven perceptual modules modeling inheritance (via coded algorithms) and experience (via statistical models).

2 Face localization in unconstrained images

Any real world face recognition system requires a reliable face localization (detection) stage that needs to be accurate and quick at the same time. Finding faces in unconstrained images presents many challenges to FR systems due to large variations of intrinsic (head pose, face expression, makeup, jewelry, etc.) as well as extrinsic (lighting, occlusions, blur, defocus, etc.) face image formation factors. To remedy these variations, we proceed by augmenting a baseline color-blind rotation-sensitive detector [11] by taking into account skin color, facial landmarks and face geometry.

2.1 Skin color mapping



(a) (b) (c)

Figure 1: (a) skin vs. (b) non-skin tone distribution and (c) resulting skin likelihood given color in RGB axes

We approach the problem of *skin mapping* by determining a real-valued skin likelihood map over any given image with a pixel-wise mapping function $s : C \rightarrow [0,1]$, where C is some color space and the skin likelihood values are real numbers in the range [0,1]. Researchers studied various color spaces [12], but for simplicity we start with RGB and use other spaces, as needed.

We compute skin and non-skin color histograms shown in Figure 1 using skin labeled data [10], [13]: (a) skin color forming a near-normal cluster in RGB, (b) non-skin color grouped around the gray-scale diagonal, and (c) conditional probability of skin (given color). The axes correspond to the color components that are quantized into 16 bins each. Each sphere has its bin's color with its size reflecting the bin's likelihood. Note that there is not much overlap between the skin and non-skin clouds, thus one can build a robust skin color classifier.

Bayesian skin mapper is based on conditional probability estimate from the source skin and non-skin pixels: $P(s|c) = P(c|s)p(s) / (P(c|s)P(s)+P(c|n)P(n))$, where $P(s|c)$ is the probability of skin given

color in (c), with P(*c*|*s*) and P(*c*|*n*) given by normalized histograms (a) and (b), P(*s*)=|*skin*|/|*all*|, and P(*n*)=1-P(*s*) with *s* = skin, and *n* = non-skin. The optimal threshold for skin/non-skin classification is ½, which is confirmed by our experiments. The method is simple and fast, as the skin mapping problem is reduced to a table look-up. It is data-driven, assumes no predefined distribution for the colors and can easily be conditioned by more data samples at any time. However, this approach may require a substantial amount of labeled data to build a general histogram for unconstrained images.

Artificial neural network (ANN) classifier is a fully connected multi-layer perceptron (MLP) that models the skin likelihood in Extended Color Space (ECS), e.g. concatenating [RGB, HSV, YCbCr], which experimentally was determined to be optimal for the task. The size of the hidden layer was set experimentally to 15. Its training involves the error back-propagation learning strategy, which converges to a certain accuracy optimum, having learning rate $\alpha = 0.02$ and momentum $\beta = 0.08$. Its generalization power to modeling of the unknown skin tone distribution is higher compared to the histogram based approach, however it requires a much longer training time.

Several metrics from information retrieval were considered:

Precision = TP / (TP+FP), Recall = TP / (TN+FN),

Fscore = 2×Precision×Recal / (Precision + Recall), Accuracy = (TP+TN) / (TP+TN+FP+FN)

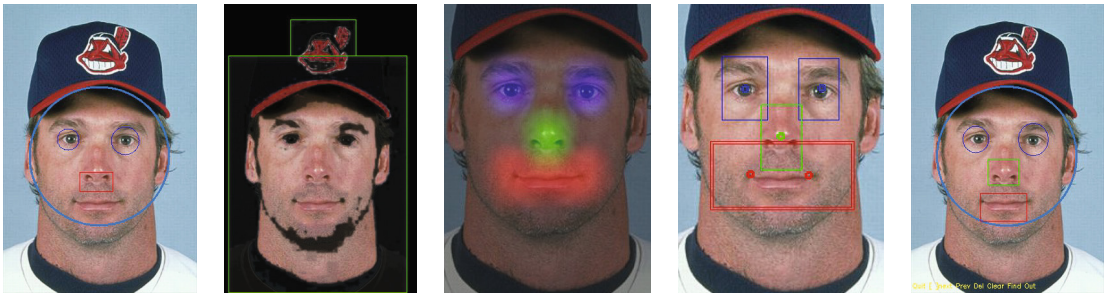
where TP = true positive, TN = true negative, FP = false positive, and FN = false negative.

skin tone mapper	recall	precision	F-score	accuracy
HIST [RGB]	0.93	0.92	**0.93**	0.86
HIST [HSV]	**0.94**	0.93	**0.93**	0.88
HIST [Lab]	0.89	**0.94**	0.92	0.84
ANN [RGB,HSV,YCbCr]	**0.94**	0.90	0.92	**0.91**

Table 1: skin mapper Bayesian (HIST) and artificial neural net (ANN) accuracy results in various color spaces

As shown in Table 1, both Bayesian (HIST) and the artificial neural net (ANN) based skin detectors performed comparably well with respect to recall, precision and F-score, but ANN-based mapper showed a greater accuracy, hence claiming a greater generalization power.

2.2 Facial landmark detection



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

Figure 2: (a) incorrect baseline, (b) skin map, (c) CNN heat-map, (d) landmarks, and (e) corrected detection

Facial landmarks detection is another important component of face localization. We employ the convolutional neural net (CNN) approach [14] and complement it with our own landmark verification stage based on encoder-decoder ANN. The color landmark mapping algorithm handles unconstrained images mapping eyes, nose and mouth blobs based on the features it learned from a collection of standard data-sets [15], [16]. The landmarks are derived from the heat maps by their major peaks through non-maxima suppression and adaptive threshold. Figure 2 shows that (a) our baseline detector mistakes nose for a mouth, then (b) our robust skin mapper narrows down the detection area,

(c) our CNN heat maps correctly overlay the landmarks, (d) landmarks are correctly localized, and (e) shows the corrected output. In all sub-figures we use red for mouth, green for nose, blue for eyes.

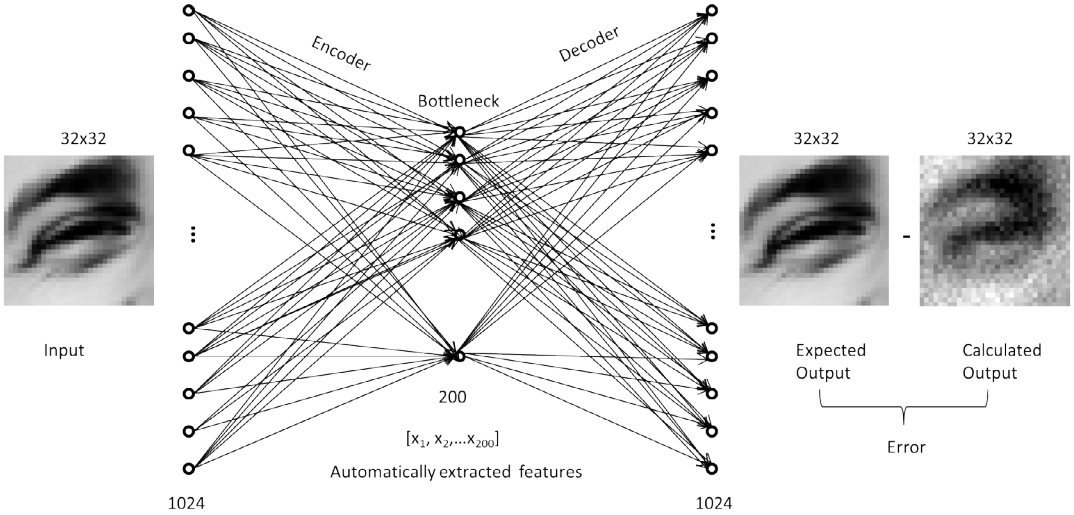


Figure 3: encoder/decoder artificial neural network assembly for landmarks verification

Our ANN-based landmark detector has a two sub-stages. The first network automatically extracts the significant features from an image patch, and the second network classifies the encoded feature vector into landmark/non-landmark, as shown in Figure 3. The advantages of this approach are: *i*) instead of guessing on statistical or structural features of an image patch, we use a basic encoder network [17] to learn the *prominent* features automatically by minimizing the image decoding error; *ii*) data dimensionality is considerably reduced; and *iii*) a complex decision mechanism based on statistical learning solves the landmarks verification problem for the source image region.

2.3 Robust face detection

The proposed face detection module is an ensemble of three agents working together: gray-scale face detector[11] complimented by the described color-aware skin mapper and the landmark detector.

method	recall	precision	F-score
Viola-Jones	0.67	0.88	0.76
Android	0.48	**0.91**	0.63
Luxand FaceSDK	0.74	0.87	0.81
FaceFinder (ours)	**0.80**	0.85	**0.82**

Table 2: face detection accuracy of different systems on FDDB benchmark

Our *skin mapping* module (run in parallel with the base Viola-Jones detector) helps diminish the non-skin regions reducing false alarms, while enhancing the large skin blobs thereby recovering missing face candidates. The color enhanced large skin blobs are then run through the color-based landmark detector, which helps identify them as face candidates that can be rectified by their eye lines and re-inspected by another instance the base face localizer for new possible faces not found originally by the gray-scale face detector. As Table 2 reveals, our FaceFinder's accuracy is on par with or better than the leading commercial and open-source face detectors we tested.

3 Single image per person (SIPP) face matching

Inspired by the human ability to match faces by a single visual sample, our SIPP approach disallows multiple samples per person, and uses a combination of key-spot [18]–[20] along with the holistic [21], [22] descriptors to ensure the overall face matching accuracy, emphasizing each descriptor's strengths, weighting them according to their individual accuracy on the available benchmark data-sets [15], [23], [24], and combining their distance functions in a generalized geometric mean [10].

dataset	CalTech [23]		ColorFERET [15]		IndianFacesDB [24]	
top-n	FaceSDK	FaceMatch	FaceSDK	FaceMatch	FaceSDK	FaceMatch
1	0.98	0.98	0.74	0.88	0.69	0.79
3	0.99	0.98	0.75	0.89	0.73	0.85
5	**0.99**	**0.99**	0.76	**0.90**	0.76	**0.87**

Table 3: Luxand FaceSDK vs. our FaceMatch hit rate accuracy in top-n queries on standard benchmark datasets

On the relatively easy CalTech faces set, accuracy figures of both contenders are high. On the more challenging NIST ColorFERET benchmark, FaceSDK clearly yields to FaceMatch. The accuracy on even more challenging IndianFacesDB dataset is noticeably lower for both competitors probably due to some extreme head pose variations, but FaceMatch still outperforms FaceSDK.

4 Conclusion

Face is arguably the most important object to human visual system to handle, hence our amazing abilities to detect and recognize them often from a single sight. Inspired by this (often taken for granted) visual functionality typical of many primates, we proposed a computational approach to face localization and matching that uses existing well performing components (as hard-wired abilities) optimizing them for the given data (emulating real-world experience), and keeping them open for change as needed (thus emulating adaptation).

Our face detection method relies on pre-trained baseline grayscale algorithm that is improved by our color-aware skin tone and landmark detection modules that are invariant to affine transformations. Some of them do require training, which sometimes can be done on-line, e.g. for conditional histograms. Our biologically inspired ANN-based classifiers are intuitive and computationally light, performing in near-real time. Our SIPP approach in FaceMatch allowed us to avoid expensive data labeling and training, yet we attain the accuracy and speed on dynamically changing web-scale datasets that is on par with or better than the existing commercial systems.

Our future R&D may involve building more modules with on-line learning capabilities including human in the loop. We plan to experiment with mobile devices (autonomously moving, wearable or hand-held) that allow gaining real-world experience and communication with humans on the go.

Acknowledgement: This work is supported by the Intramural Research Program of the U.S. National Library of Medicine (NLM), National Institutes of Health. We also thank Zhirong Li and Lan Le for their help with data sets and experiments.

References

[1] L. Chang and D. Y. Tsao, “The Code for Facial Identity in the Primate Brain,” *Cell*, vol. 169, no. 6, p. 1013–1028.e14, 2017.

[2] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.
[3] C. Lu and X. Tang, "Surpassing Human-Level Face Verification Performance on LFW with GaussianFace," *CoRR*, vol. abs/1404.3840, 2014.
[4] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An All-In-One Convolutional Neural Network for Face Analysis," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2017.
[5] J. R. Beveridge *et al.*, "The challenge of face recognition from digital point-and-shoot cameras," in *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, 2013, pp. 1–8.
[6] G. B. Huang and E. Learned-Miller, "Labeled Faces in the Wild: updates and new reporting procedures," UMass, UM-CS-2014-003, 2014.
[7] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1931–1939.
[8] F. Rosenblatt, "Neurocomputing: Foundations of Research," J. A. Anderson and E. Rosenfeld, Eds. Cambridge, MA, USA: MIT Press, 1988, pp. 89–114.
[9] F. Crick, "The recent excitement about neural networks," *Nature*, vol. 337, no. 6203, p. 129—132, Jan. 1989.
[10] E. Borovikov and S. Vajda, "FaceMatch: real-world face image retrieval," in *Recent Trends in Image Processing and Pattern Recognition*, 2016.
[11] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, pp. 137–154, 2004.
[12] D. Malacara, *Color vision and colorimetry: theory and application*. SPIE Press, 2002.
[13] M. Jones and J. M. Rehg, "Statistical Color Models with Application to Skin Detection," in *International Journal of Computer Vision*, 2002, pp. 274–280.
[14] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From Facial Parts Responses to Face Detection: A Deep Learning Approach," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, Washington, DC, USA, 2015, pp. 3676–3684.
[15] "The Color FERET Database." [Online]. Available: http://www.nist.gov/itl/iad/ig/colorferet.cfm.
[16] M. Köstinger, 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 *ICCV Workshops*, 2011, pp. 2144–2151.
[17] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[18] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[19] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *European Conference on Computer Vision*, 2006, pp. 404–417.
[20] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *IEEE International Conference on Computer Vision*, 2011, pp. 2564–2571.
[21] C. E. Jacobs, A. Finkelstein, and D. H. Salesin, "Fast multiresolution image querying," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, New York, NY, USA, 1995, pp. 277–286.
[22] A. Satpathy, X. Jiang, and H.-L. Eng, "LBP-Based Edge-Texture Features for Object Recognition," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1953–1964, May 2014.
[23] CalTech, "Caltech Frontal Face Dataset." [Online]. Available: vision.caltech.edu/archive.html.
[24] V. Jain and A. Mukherjee, "The Indian Face Database," 2002. [Online]. Available: http://vis-www.cs.umass.edu/~vidit/IndianFaceDatabase/.