Face Match for Family Reunification: Real-world Face Image Retrieval

Eugene Borovikov*, Michael Gill
U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA
(FaceMatch@NIH.gov)
Szilárd Vajda
Central Washington University, 400 E. University Way, Ellensburg, WA 98926, USA
(Szilard.Vajda@cwu.edu)

ABSTRACT

Despite the many advances in face recognition technology, practical face detection and matching for unconstrained images remain challenging. A real-world Face Image Retrieval (FIR) system is described in this paper. It is based on an optimally weighted image descriptor ensemble utilized in a single-image-per-person (SIPP) approach that works with large unconstrained digital photo collections. The described visual search can be deployed in many applications, e.g. person location in post-disaster scenarios, helping families reunite quicker. It provides efficient means for face detection, matching and annotation, working with images of variable quality, requiring no time-consuming training, yet showing commercial performance levels.

Keywords: Face Detection, Face Recognition, Image Retrieval, Family Reunification

INTRODUCTION

The Content Based Image Retrieval (CBIR) technology has seen significant advances recently resulting in many useful web-scale image search techniques (Dharani & Aroquiaraj, 2013). Several web search engines (e.g. bing.com/images, images.google.com, yandex.com/images) employ those techniques to provide visual search capabilities. The face recognition (FR) technology has also seen a considerable progress during the last decade, in several cases approaching human-level accuracy in face detection and verification tasks (Naruniec, 2010; Tan, Chen, Zhou, & Zhang, 2006; Zhang & Zhang, 2010), especially in well-controlled environments such as studios.

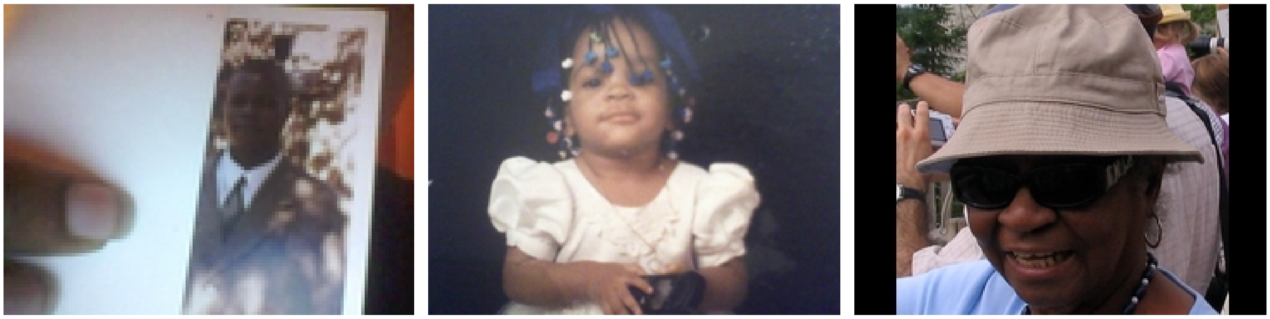


Figure 1: Unconstrained images present challenges to face recognition systems.

Modern web-based FR solutions (e.g. in facebook.com or plus.google.com) may work well with limited face datasets (e.g. user circles, family albums) that tend to contain tagged pictures of the same few individuals (e.g. family and friends) with multiple photos per person, which allows for user-specific recognition model training. Our experience did not provide us with an abundance of publicly available *single image per person* (SIPP) face image retrieval systems that can work effectively using *no training* with millions of unconstrained face images, presenting many challenges for such systems in practice, e.g. disaster recovery:

- no constraints on gallery or query pictures, as in Figure 1
- often suboptimal quality images for query and gallery,
- dataset size: web-scale collections with many near-duplicates[1]
- large inconsistency in query and gallery face appearance.

Many of those challenges are being addressed by modern FR systems thanks to the emergence of labeled datasets with constrained-free images (Beveridge et al., 2013; Huang, Ramesh, Berg, & Learned-Miller, 2007; Kemelmacher-Shlizerman, Seitz, Miller, & Brossard, 2016) utilized for various competitions. Development of such challenging datasets presents a great opportunity to assess capabilities of the existing systems on the real-world data, and then improve them or develop some new capabilities, ultimately approaching a human-level visual matching accuracy.

Typical FR systems would approach the face recognition problem in one of the two formulations (Zhou et al., 2014): *verification* (answering if photos depict the same person) or *identification* (suggest the person ID by visual similarity to the query image). Such systems usually require some sort of model training, using multiple photos per individual. They would typically work with a set of visual features extracted from images by learning a measure of visual similarity, modeling human visual perception of faces. While modern automatic face classification and verification methods can work fairly well on good quality (fairly well lit, sharp, 80×80 pixels or better) face images, their performance degrades quite rapidly as the image quality drops (e.g. due to blurring, scaling, re-compression, etc.) causing significant degeneration of the visual attributes (Scheirer, Kumar, Iyer, Belhumeur, & Boult, 2013) they rely on.

We approach our face matching task as a *face image retrieval* (FIR) problem: given a query image, we aim at returning visually similar faces from a *dynamically changing* photo gallery, thus effectively reducing the search space from several thousands to about 20 likely candidates conveniently displayed on a single page. Our open-set approach practically out-rules person-specific training and uses the accuracy evaluation methods (e.g. top-N hit rate) that are typical of CBIR, rather than those typical of FR, e.g. Receiver Operating Characteristic (ROC), although the latter could also be used for compatibility reasons (Fawcett, 2006).

Our SIPP face image retrieval methodology has been deployed in a real-world *face retrieval* system (FaceMatch), detecting and matching faces in arbitrary orientation or lighting conditions. Handling large scale photo collections, our system requires no training while dealing with any open sets of images. FaceMatch R&D project attempts to solve most of the challenges mentioned, providing the following functionality:

- semi-automatic annotation for faces and landmarks
- accurate face detection robust to scale and rotation
- image descriptors ensemble for improved face image match

We present FaceMatch evaluation results for the face detection, matching and retrieval tasks, using several publicly available data sets, some of which were annotated in our laboratory. Our face image retrieval system is naturally fine-tuned for face detection and matching, but it can also be used for general-purpose object and scene matching, thus providing a rich set of tools for practical large-scale image collection exploration and manipulation.

[1] visually almost identical, but not the same

In what follows, we discuss our image repository, detail on the proposed methodology and present major components of our FaceMatch (FM) system. In each section, we review the relevant publications, describe our approach, and present experimental results.

IMAGE DATA COLLECTIONS

Our approach to face image retrieval is driven by data. The described methodology automatically extracts and weights image features based on statistics. We build and utilize annotated image repositories that provide us with ground truth (GT) for the accuracy evaluation and optimization of individual components, e.g. skin mapping for face localization.

Image annotation for *face detection* typically consists of localized face regions (and optionally face landmarks: eyes, nose, mouth, and ears), optional gender and age groups, and some skin patches. Such annotations are done semi-automatically, providing the human annotator with initial face and landmark localization, which can be manually corrected or completed.

Ground truth for *face matching and retrieval* involves labeling face images with person ID[2] that are used to assess the quality of retrieval accuracy. Our system targets unconstrained image datasets, e.g. photos from natural disaster events collected by People Locator (PL). PL dataset consists of 40 thousand weakly text-labeled mostly color, low quality images, some of which are shown in Figure 1. PL image repository is changing over time, as disasters happen (Thoma, Antani, Gill, Pearson, & Neve, 2012).

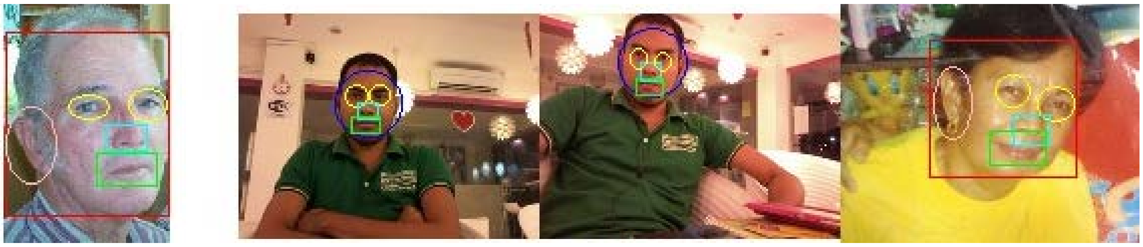


Figure 2: Face and landmarks annotation examples

To better organize PL repository, we have developed several cross-platform image processing and annotation tools to

- reduce data size by removing near-duplicates,
- outline faces and profiles as rectangular regions,
- localize facial landmarks: eyes, nose, mouth, and ears,
- extract skin patches from the skin-exposed regions.

These tools are used to partially annotate various image collections with the correct face and profile locations and facial landmarks, as shown in Figure 2. Our near-duplicate detector is based on an efficient image retrieval by sketch (Jacobs, Finkelstein, & Salesin, 1995) method using color wavelets, and capable of comparing millions of image signatures in a second. Our annotation tools are semi-automatic because image processing is never perfect, and each dataset has some unique characteristics, requiring a human annotator to confirm the annotation as ground truth, which is then used for accuracy assessments and methodology improvements. Face and landmark annotation tools use the corresponding capabilities of our FaceMatch library, to be

[2] unique alpha-numerical sequence, *not* revealing the true person identity

described later in more detail. Using these web-based and desktop annotation tools, our team provided image ground truth for several thousand PL images, producing several annotated sets:

PL-Faces: consists of 2882 low resolution, color PL images, with ¾ of face regions being frontal and about ¼ being profile views. The average face and profile diameters are 40 and 50 pixels respectively.

HEPL-500: is a subset of PL containing 500 images from 2011 Haiti earthquake, containing a large variety of faces. Some of them are over-exposed, blurry or occluded as shown in Figure 1.

Boston-2013: image set consists of 417 low resolution images reported in connection to the 2013 Boston marathon bombing, with 28 photos of the two main suspects.

PL-Skin: was created for the skin color mapping needs. 7680 PL images producing 33,431 patches from faces, arms, legs were annotated resulting in a total of 13 million pixels, of which 7 million were labeled as *skin* and 6 million were labeled as *non-skin*.

The images were selected to include a large range of skin tones, environments, cameras, resolutions, and lighting conditions. Some of the images contain multiple human subjects. The quality of the images varies significantly in illumination, resolution and sharpness.

Some of the datasets were created in *collaboration* with other research labs, benefiting face recognition research community in general with some additional annotations and checks:

Lehigh Faces: set was obtained through our collaboration with Lehigh University (Kim, Huang, & Heflin, 2011) exhibiting wide variations in background and pose, with mostly light skin tones. C1 subset contains 512 unconstrained but near frontal looking images of celebrities. C2 subset contains 550 images, but with a greater variety in face appearances.

Compaq Skin: set (Jones & Rehg, 2002) of nearly 1 billion skin/non-skin labeled pixels for training/testing skin tone classifiers. We had to filter out some obvious (black or white) outliers, thus reducing the dataset by about a million points.

Additional meta-data annotation (e.g. skin tone, age group, gender) have also been introduced for most datasets. The annotated repository is used for improving face detection and matching performance.

We also utilize some publicly available *benchmark* datasets depicting humans in unconstrained environments for algorithm evaluation and tuning:

CalTech Faces: set consists of 450 frontal views of 29 subjects, which are taken under varying lighting and background conditions.

Indian Faces: set contains 676 face images of 61 individuals, photographed in a studio, exhibiting large variations in head pose, face expression, and lighting.

ColorFERET: set contains 2413 facial images of 856 individuals showing frontal and left/right profile head pose variation, optional glasses, and various facial expressions.

FDDB: Face Detection Data Set and Benchmark contains 2845 images with 5171 unconstrained faces (V. Jain & Learned-Miller, 2010).

For some of the mentioned sets (e.g. CalTech and Indian Faces) we have contributed landmark annotations in addition to the supplied head/face regions. Our experiments use those sets to test FaceMatch performance, and the evaluation results are presented in the respective sections.

COLOR-AWARE FACE AND LANDMARK DETECTION

Reliable face localization is the first critical step in any face matching application. Color-blind face detection has been well researched and some efficient detectors have been developed (Zhang & Zhang, 2010). Some of those detectors can run in near-real time (Viola & Jones, 2004), but they typically come with pre-trained models that may work well for the near-frontal views of faces, but fail on many unconstrained images where faces could be arbitrarily positioned, occluded, blurred or sub-optimally lit, as in Figure 1.

To improve on this base-line face detection accuracy, we propose to use color as one the most important cues for face (and its landmarks) presence or absence (Deng & Pei, 2008; Hsu, Abdel-Mottaleb, & Jain, 2002). To have the resulting color-aware detector run at similar near-real-time speeds as the base, we propose to utilize graphics processing units (GPU) for gray-scale face detection, skin color mapper and basic image processing, while running higher level components on separate CPU cores using multi-threading techniques, thus taking advantage of CPU/GPU parallelism.

Our face localization sub-module (FaceFinder) is an ensemble of several algorithms working together: base grayscale *face detector* (Viola & Jones, 2004), and color-aware neural network based *skin mapper* with *landmark detector* (developed by us), as shown in Figure 3: our additional modules help recover missing faces while reducing the number of false alarms.

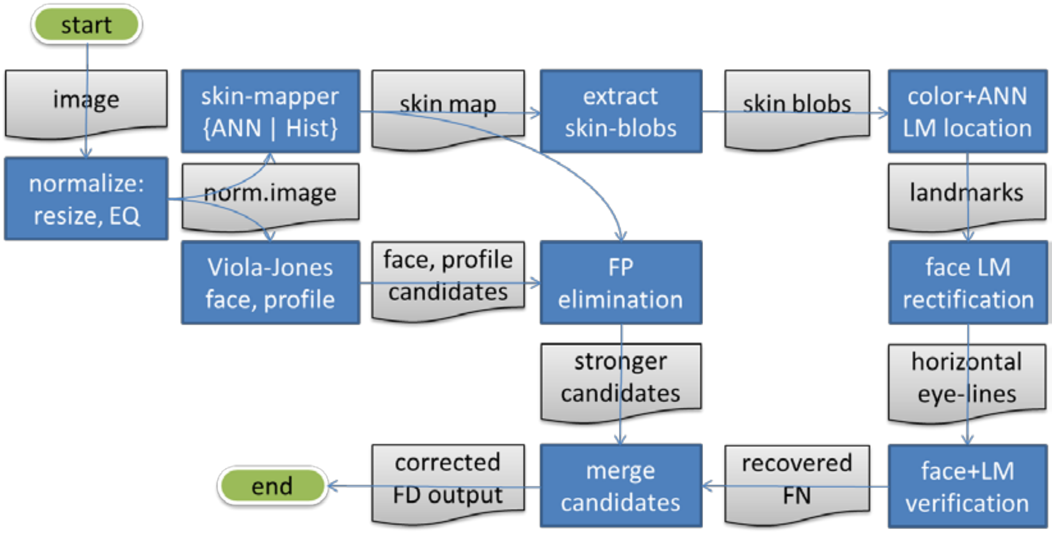


Figure 3: FaceFinder components (blue) with data items (gray) and parallel execution flow

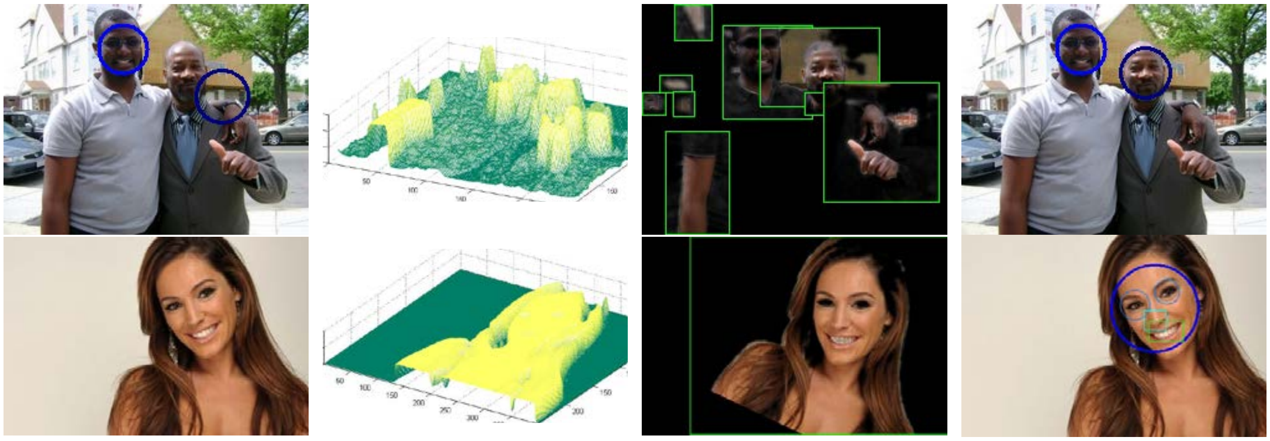
Our *real-valued skin mapping* module (run in parallel with the base detector) helps diminish the non-skin regions (reducing some false alarms) and enhance the large skin blobs (recovering some missing face candidates). The color enhanced large skin blobs are then run through the color-based landmark (eyes, mouth) detector (Hsu et al., 2002), 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 grayscale face detector.

Some visual results of our FaceFinder major stages are presented in Figure 4: base detection (a) gets corrected by computing the real-valued skin color map (b), which is used to remove the false detection and recover the missing candidates by landmark localization and eye-line rectification (c) to produce the output (d).

To evaluate the face detection accuracy of FaceFinder, we have considered a variety of image collections, described in the section describing our image collections. For evaluation, we have used the traditional information retrieval metrics: R=Recall, P=Precision and F=F-score, defined as

$$R = M/S,\ P = M/D,\ F = 2PR/(P+R),$$

where respective counts are M=Match, S=Source, and D=Destination. Match was incremented every time there was a considerable overlap between the detected and the source region, known from the pre-annotated ground-truth data.



(a) base detection (b) skin color map (c) rotated skin blobs (d) corrected output

Figure 4: Corrected face detection using real-valued skin map and landmarks

The overlap of two rectangular face regions *A* and *B* can be calculated as

$$L = \frac{|A \cap B|}{|A \cup B|}$$

and if the overlap $L > 0.5$, we considered that to be a correct detection (match).

Table 1: Face detection accuracy on different datasets

Data	Methods	Recall	Precision	F-score
HEPL-500	VJFD	0.76	0.87	0.81
	FaceFinder	**0.77**	0.89	**0.83**
	CmrMbl	0.68	0.87	0.76
	OpnMbl	0.47	**0.94**	0.62
	FaceSDK	0.75	0.91	0.82
	OpnSrc	0.33	0.92	0.49
FDDB	VJFD	0.67	0.88	0.76
	FaceFinder	0.75	0.86	**0.81**
	CmrMbl	0.63	0.76	0.69
	OpnMbl	0.48	**0.91**	0.63
	FaceSDK	**0.76**	0.87	**0.81**
	OpnSrc	0.61	0.79	0.69
Lehigh-C1	VJFD	0.95	0.81	0.88
	FaceFinder	**0.97**	0.91	**0.94**
	CmrMbl	0.95	0.92	**0.94**
	OpnMbl	0.52	**0.93**	0.67
	FaceSDK	0.96	**0.93**	**0.94**
	OpnSrc	0.83	0.91	0.87

Table 1 summarizes the accuracy figures for the following face detection systems that our FaceFinder is compared with:

VJFD: baseline open-source system (Viola & Jones, 2004)
FaceFinder: our skin-tone based and landmark-aware detector
CmrMbl: commercial mobile face detector
OpnMbl: open-source mobile face detector
FaceSDK: commercial desk-top based face detector
OpnSrc: desk-top based open-source detector OpnSrc (Zhu & Ramanan, 2012)

Overall, our FaceFinder accuracy is on par with, if not better than, those of the leading commercial and open-source face detectors. The benefits of our face detection subsystem can be summarized as follows:

- recovering faces missed by baseline detection stage,
- overruling false alarms by eliminating low skin regions,
- detecting rotated faces by recovering their landmarks.

Our color-aware face detection method is robust to the affine transformations, lighting and image noise. Provided enough CPU/GPU power, our adaptive cross-platform multi-core implementation is more accurate than its baseline face detector, yet it runs on average at about the same speed as the CPU-based VJFD implementation on the same hardware.

The described FaceFinder functionality is utilized in the subsequent color-aware face image ingest and retrieval stages, whenever reliable face detection is needed, e.g. during ingest or query requests.

FACE IMAGE RETRIEVAL

Our FaceMatch (FM) method addresses the *single image per person* (SIPP) face image *retrieval* problem that is optimized for interactive-time visual queries in large *dynamic* collections of face pictures in unconstrained environments, e.g. arbitrary resolution, scale, and illumination. Thus our approach is different from face *verification* (1:1, as our decision is not binary) or face *identification* (1:*N*, as our image sets are dynamic) that are typically addressed by FR systems.

Although in the last couple of decades the FR community has considerably advanced the field and produced a large number of strong methods, FR in general conditions remains to be an open problem that's being researched actively (Azeem, Sharif, Raza, & Murtaza, 2014; Jafri & Arabnia, 2009; Sharif, Mohsin, & Javed, 2012). Beham (Beham & Roomi, 2013) gives a good overview of FR techniques and divides them in the following major groups (holistic, feature-based, and soft-computing), providing normalized accuracy (NA) figures, pointing out their advantages and drawbacks.

Unconstrained, SIPP face retrieval from a large, dynamically changing (open-set) reference gallery basically requires its face matching to be training-less, robust to pose, occlusion, expression, lighting, and fast, i.e. essentially modeling human perception of unfamiliar faces from a single photo and utilizing some fast approximate indexing for efficiency.

Several very promising methods (Gao & Qi, 2005; A. K. Jain, Klare, & Park, 2012; Tan et al., 2006) have been proposed over the past decade, and more recent papers describe systems that are comparable to the human performance at face verification task (Lu & Tang, 2014; Taigman, Yang, Ranzato, & Wolf, 2014). This kind of accuracy typically implies (deep) learning systems with a substantial training stage using hundreds or thousands shots per person, and their matching time may still be not very practical for large scale interactive searches.

The one-shot similarity kernel (Wolf, Hassner, & Taigman, 2009) approach to face matching uses a special similarity measure to produce some impressive face matching results on Labeled Faces in the Wild (LFW) collection (Huang et al., 2007). We cannot utilize this approach directly, as it requires some training with the background examples.

Face matching

We propose a scalable visual search method addressing the face image retrieval problem for dynamically changing image collections. Face image queries can be executed after all face regions in the image collection are detected and their image descriptors are indexed. The proposed method uses an ensemble of image descriptors that capture various important aspects of a face and thus accommodates wide variations in face appearance mentioned in the introduction. We have experimented with several individual image descriptors for face matching:

HAAR: our modification of the holistic color-aware Haar wavelet (Jacobs et al., 1995) descriptor with three color-band structure, where each band keeps (a) the average of 128×128 bins, and (b) signed integer offsets of the 40 largest wavelet coefficients. This accelerates the search and reduces the storage for the visual index.

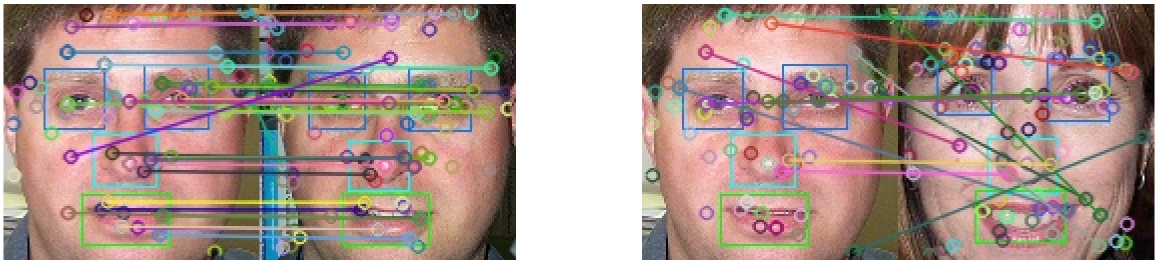
LBPH: our adaptation of the holistic gray-scale Local Binary Pattern Histogram descriptor (Ahonen, Hadid, & Pietikäinen, 2004) using histogram of 256 local (radius=1) binary patterns collected from 8×8 cell grid imposed on the input face region, resulting in 214=16384 float values signature. The descriptor is robust to lighting showing good frontal face matching performance.

ORB: Oriented FAST (Rosten, Porter, & Drummond, 2010) and Rotated BRIEF (Calonder, Lepetit, Strecha, & Fua, 2010) key-point descriptor (Rublee, Rabaud, Konolige, & Bradski, 2011) that computes 500 binary feature values taking 64 bytes for each key-point, measuring pixel intensity differences at random image locations within the key-spot region, and recording 1 for a positive difference, and 0 for a negative one. ORB key-points tend to cluster around corner-rich face features (eyes, mouth, nose, ears) and using fast matching procedure for the descriptors.

SIFT: gray-scale Scale Invariant Feature Transform (Lowe, 2004): computes 128 float value signature for each key-point location, which is characterized by the texture that is robust to object scale, translation and rotation within an image. This provides for a great flexibility with respect to the head pose and the distance to the camera.

SURF: gray-scale Speeded Up Robust Features (Bay, Tuytelaars, & Gool, 2006), as a quicker alternative to SIFT, it computes 64 float value signature for each key-point location that is also characterized by the texture that is robust to object scale, translation and rotation within an image. SURF computation is accelerated by the use of the integral images and a discrete approximation for the Hessian matrix.

RSILC: our Rotation and Scale Invariant Line Color (Candemir, Borovikov, Santosh, Antani, & Thoma, 2015) descriptor using 3 color band structure producing 512 float values per band per key-line. In addition to the key-line local gradient and color information, this descriptor also captures the spatial information, e.g. what other key-lines are visible at which angles from the given key-line. Thus RSILC is a larger and more accurate image descriptor then SIFT or SURF, but it is significantly slower to compute and compare than its key-point competitors by an order of magnitude, which suggests its acceleration via GPU.



(a) same person, different faces (b) different people

Figure 5: SIFT based matching performance of the system on two example faces

Given a query face image, the goal is to match its descriptors against the index of the existing face descriptors, and output a list of likely face candidates ordered by similarity. The proposed matching technique does *not* assume that many faces of the same subject are present in the database, and it is robust to illumination, scale and affine transformations.

Figure 5 presents two unrestricted key-spot matching examples using SIFT descriptors. The left pair shows matches between two different photos of the same person: the number of correctly matched locations is relatively high. The right pair shows the faces that belong to different people: there are evidently fewer sensible matching locations, e.g. note the non-matched key-spots at the chin location of the faces. Experiments on several datasets revealed that

- single descriptor is insufficient for accurate retrieval,
- some key-spot matches need to be filtered as outliers,
- face landmarks help filter and weigh the matches.

Having several image descriptors per face (HAAR, LBPH, SIFT, etc.) allowed us to capture both holistic and key-spot information about each face, improving the overall matching power by leveraging the strengths while downplaying the weaknesses of the descriptors. We experimented with similarity distance-based and similarity rank-based feature combination strategies. The combinations used individual distances $d_i \in [0,1]$ (or ranks) and descriptor matching confidence weights $w_i \in [0,1]$:

DIST: *weighted distance product* $d = \prod d_i^{w_i}$

RANK: *rank-based* combination based on Borda

The weighted descriptor ensemble hence allows:

- combination of holistic with the key-spot based image descriptors,
- utilization of color along with the texture information,

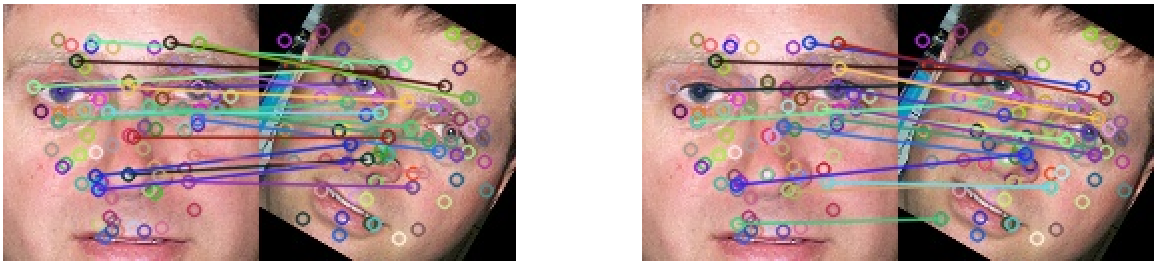
The optimization procedures are performed using the non-linear simplex (Nelder & Mead, 1965) method maximizing the retrieval accuracy expressed as F-score or *hit rate*, i.e. the frequency of retrieving the correct subject given a probe photo in a top-*N* query, i.e. for a set of query images Q, define the hit rate for top-*N* matches as

$$HitRate(N) = HitCount(N, \mathrm{Q})/|\,\mathrm{Q}|$$

where *HitCount* counts the successful top-*N* matches using the query set of size |Q|.

Enhancing key-spot matching accuracy and speed

As Figure 5 suggests, there may be some key-spot mismatches that may in turn cause some false hits in face image queries. To improve matching confidence, our key-spot descriptor matching scheme includes the descriptor symmetric match *cross-check* to ensure that best match relationship works both ways. Our key-spot matching algorithm ignores descriptor matches whose distance is greater than two minimum distances across the matching pool, but it still may result in some false hits.



(a) RANSAC scale and rotation

(b) MEADOW scale and rotation

Figure 6: Spurious SURF key-spot match filtering to ensure geometric consistency

To further improve the key-spot descriptor matching accuracy, we filter out the outliers among the two-way descriptor matches via the inter-view homography (Chum, Pajdla, & Sturm, 2005) based RANdom SAmple Consensus (RANSAC) algorithm (Zuliani, Kenney, & Manjunath, 2005). This iterative statistical method computes and uses an affine transform between two images (homography) of the same (or similar) object to assert the key-spot consensus. It works quite well for the near-frontal views of in-plane rotated and scaled faces, as shown in Figure 6, but it may slow down the face matching process because of is iterative nature and having to estimate the homography matrix at each iteration.

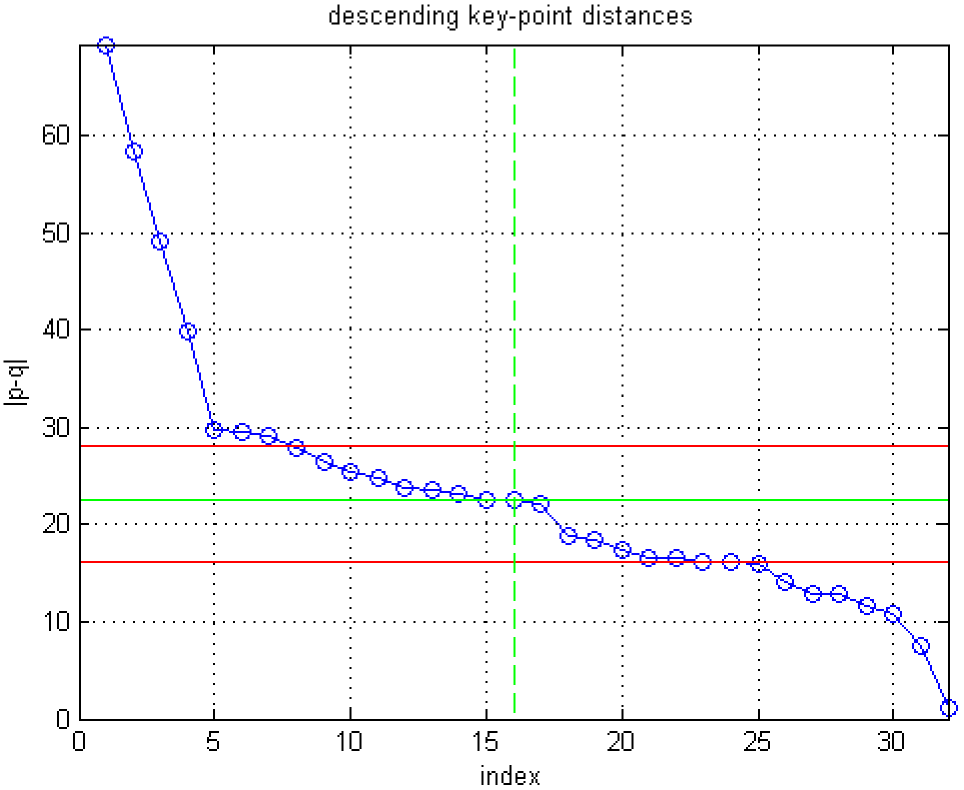


Figure 7: MEADOW filters distance outliers above and below the median deviation lines (red) with respect to the sample's median (green). The vertical axis is the geometric distance between key-points p and q, while horizontal axis is the index of the distance.

As a quicker alternative to RANSAC, we researched and developed MEdian-based Anomalous Distance Outliers Weeding (MEADOW) method. As the name suggests, the method weeds out the key-spot outliers, i.e. matches with too unlikely geometric distances between the corresponding key-spots. Compared to RANSAC, MEADOW is intended to be

- more efficient: no iterative estimation of homography
- less constrained: no key-spot co-planarity assumption

MEADOW is expected to be less accurate than RANSAC in general, but for practical face image matching applications, their accuracies are comparable.

For each two-way descriptor match MEADOW computes the Euclidean distance between their key-points (not descriptors) p and q, and we discard that match as a false positive, if that distance $D = |p\text{-}q|$ is an outlier among all the distances in the match sample: $|D\text{-}M| > T$, as shown in Figure 7, where M is the sample's distance median (dashed green lines), and T is computed as a median deviation from M. MEADOW is a simpler (than RANSAC) method for filtering out the largest outliers from a sample, which is what we intend for the key-spot distances to ensure the key-spot geometric consistency. As we can see in Figure 6, MEADOW efficiently handles the outliers, filtering out most of the false matches, typically five times faster than RANSAC, resulting in a similar matching accuracy.

Descriptor search space partitioning

While dealing with large unconstrained face image datasets (over 40K images), our system, to be practical, needs to retrieve face images within interactive (about 1 second) turn-around time intervals. To accomplish that we researched and developed the *attribute bucketing* strategy and utilized the *approximate nearest neighbor* (FLANN) searches (Muja & Lowe, 2009).

We have noticed that our images typically carry gender and age-group meta-information, which allowed us to partition the search space into a number of age and gender groups (called buckets), which we could query in parallel using multi-threading. This allowed us to optimize our query turn-around times by a factor of 9 or more, especially when we introduced sub-bucketing within groups.

Utilization of FLANN resulted in the additional (five-fold on average) queries speed-up with a small penalty (a couple of percentage points) to the retrieval accuracy and a small one-time clustering overhead during the index loading and incremental update. Overall, the face image query turn-around times are kept under a second for our image data-sets. Provided enough multi-core processing power it should be scalable to the web-scale sets of millions of images.

Experiments

Due to the sources of our target image collections, we very rarely have more than one picture of the same person. Hence, in our face retrieval evaluations, we had to rely on a mixture of datasets, e.g. the CalTech Faces data mixed with some typical PL photos.

Table 2: Color-aware face matching top-1 hit rates

	IndianFaces		ColorFERET	
descriptor	alone	+ color wavelet	alone	+ color wavelet
color wavelet	0.52	0.52	0.78	0.78
SIFT	0.61	0.66	0.91	0.95
SURF	0.75	0.78	0.96	0.98
SURF+SIFT	0.76	0.79	0.97	0.98

For the color-aware face matching experiments, we considered IndianFacesDB and ColorFERET datasets, containing color images of male and female faces with good variations in lighting, pose, and expression.

As shown in Table 2 our color wavelet (CW) descriptor alone is a weaker matcher than any of the key-point based descriptors, but it considerably improves the query hit rates, when included in the ensemble with the stronger (but color-blind) descriptors. This behavior suggests that bringing color-awareness to the descriptor ensemble helps improve the face matching performance on color images.

Table 3: FaceMatch (FM) vs. FaceSDK (FSDK) hit rate accuracy in top-N queries

	CalTech		ColorFERET		IndianFacesDB	
top-N	FaceMatch	FaceMatch	FaceMatch	FaceSDK	FaceMatch	FaceSDK
1	0.98	0.98	0.93	0.74	0.79	0.69
3	0.98	0.98	0.96	0.75	0.85	0.73
5	0.99	0.99	0.96	0.75	0.87	0.76
10	0.99	0.99	0.97	0.76	0.90	0.79
20	**1.00**	**1.00**	**0.98**	0.76	**0.92**	0.83

For the FaceMatch overall visual feature ensemble (with optimally weighted descriptors), the top-N hit rate accuracy results on the available benchmark datasets are summarized in Table 3 in comparison with the commercial face matching engine FaceSDK: on a relatively easy CalTech dataset (with large, mostly frontal faces), accuracy figures of both FaceMatch and FaceSDK are predictably high and close to each other.

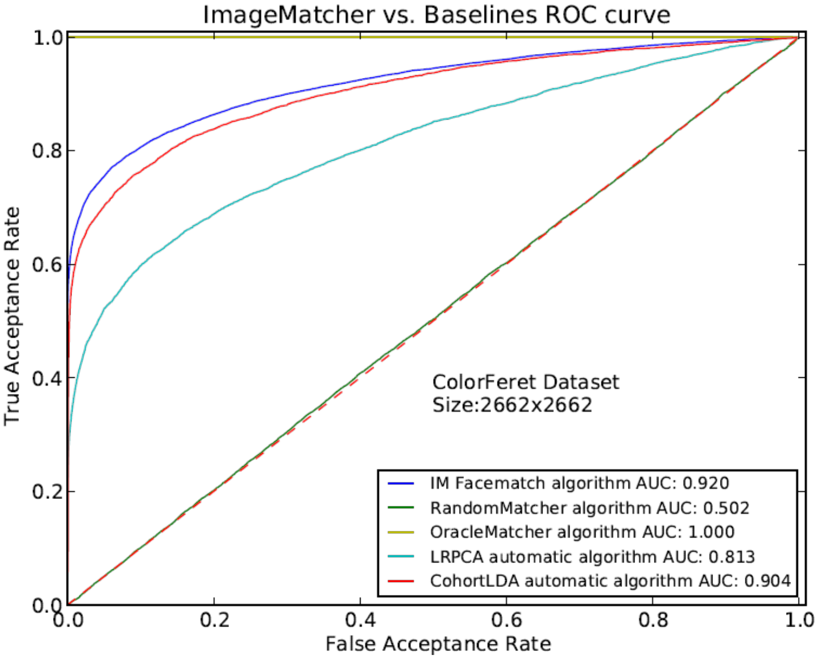


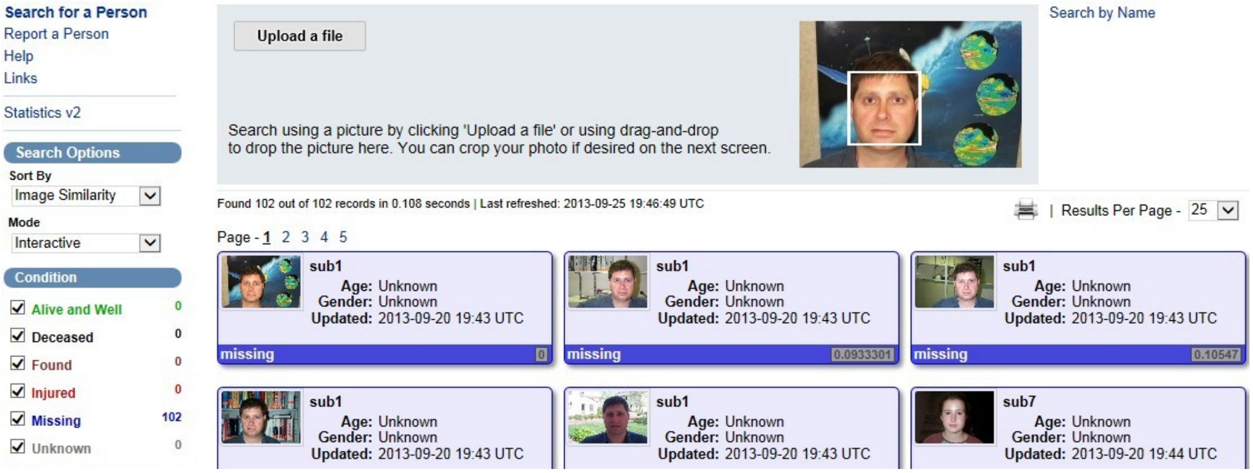
Figure 8: ROC curves for FaceMatch and baseline algorithms

On the more challenging (than CalTech) ColorFERET benchmark dataset with considerable variations in head pose and lighting, FaceSDK clearly yields to FaceMatch, which performs just as well as it does on CalTech, reaching the statistically guaranteed retrieval of the correct person within top 20 retrieved records. The accuracy on even more challenging (than CalTech or ColorFERET) IndianFacesDB dataset is noticeably lower for both competitors probably due to some extreme head pose variations, but FaceMatch clearly outperforms FaceSDK, providing the 92% likelihood of retrieving the right person in top twenty visual query results.

To perform Receiver Operating Characteristic (ROC) analysis (Fawcett, 2006) of FaceMatch in comparison to some baseline algorithms, we used the evaluation protocol for Point-and-Shoot Camera challenge posted by NIST (Beveridge et al., 2013). As Figure 8 shows, our IM FaceMatch algorithm produces a rather smooth ROC curve, and by the Area Under the Curve (AUC) measure, it outperforms the baseline LRPCA (Phillips et al., 2011) and CohortLDA (Lui, Bolme, Phillips, Beveridge, & Draper, 2012) algorithms on the ColorFERET dataset. The RandomMatcher and OracleMatcher curves are shown for the reference to the worst and the best possible matcher performance.

SYSTEM

Our production-level system is cross-platform and data-driven. The core FaceMatch (FM) library is written in portable C++11, relying on open source libraries, e.g. STL, OpenCV, and OpenMP. It is deployable for desktop applications or over as web services. The main focus for the web integration was to ensure best performance across principal FaceMatch operations, e.g. list, ingest, query and remove. Our design takes advantage of multi-core architectures by exploiting task level and functional parallelism inside all critical modules. For instance, the web service can answer multiple queries while ingesting or removing descriptors.



Search using a picture by clicking 'Upload a file' or using drag-and-drop to drop the picture here. You can crop your photo if desired on the next screen.

Sort By: Image Similarity

Mode: Interactive

☑ Alive and Well 0
☑ Deceased 0
☑ Found 0
☑ Injured 0
☑ Missing 102
☑ Unknown 0

Found 102 out of 102 records in 0.108 seconds | Last refreshed: 2013-09-25 19:46:49 UTC

| Results Per Page - 25

sub1
Age: Unknown
Gender: Unknown
Updated: 2013-09-20 19:43 UTC
missing 0

sub1
Age: Unknown
Gender: Unknown
Updated: 2013-09-20 19:43 UTC
missing 0.0933301

sub1
Age: Unknown
Gender: Unknown
Updated: 2013-09-20 19:43 UTC
missing 0.10547

sub1
Age: Unknown
Gender: Unknown
Updated: 2013-09-20 19:43 UTC

sub1
Age: Unknown
Gender: Unknown
Updated: 2013-09-20 19:43 UTC

sub7
Age: Unknown
Gender: Unknown
Updated: 2013-09-20 19:44 UTC

Figure 9: FaceMatch sample visual query results on the CalTech+PL data

The FaceMatch (FM) services are currently utilized in a real-world family reunification system, adding a visual search capability to the otherwise text based searches. The uses can inspect the details of the retrieved records and optionally re-submit queries using the retrieved faces as examples. The output of the FaceMatch module can be optionally fused with the text query results for an increased query accuracy. A sample visual query results are shown in Figure 9, and we can see how the system retrieves the faces similar to the query in the ascending distance-to-query order, observing the same person photos being at the top of the result set.

CONCLUSION

Targeting a practical system handling web-scale photo collections with real-world images, we researched and developed a *single-image-per-person* (SIPP) query-by-photo methodology (FaceMatch) working with unconstrained images of variable quality, implemented it as a cross-platform software library, exposing its face localization and image retrieval functionality via web-services, which are consumed by real-world applications, such as efficient photo collection search for the disasters management.

With real-world collections of hundreds of thousands records, FaceMatch can help reduce the user browsing set of most likely candidates to about 20, running queries at user-friendly turn-around times of about a second. FaceMatch has shown certain robustness in *cross-ethnicity* face queries, retrieving visually similar *other-ethnicity* photos faster and at times more reliably than a hospital worker could under the stress of an emergency. This could help save time and effort for the disaster event managers and for people who search for their missing relatives.

We evaluated a few state-of-the-art systems on available datasets, researched and developed methodology for face image retrieval, resulting in a software library for: (i) image near-duplicate detection, (ii) general image queries, (iii) robust face detection, (iv) efficient face matching. The major features that make FaceMatch practical for the real-world face image retrieval:

- *unconstrained* images handling,
- *training-less* single-image-per-person (SIPP) approach,
- *cross-platform* approach to the implementation.

Our technology matches the performance of the leading open-source and commercial solutions. We have made several important improvements to the existing methods and developed some new ones:

Face detection was improved by using human skin tone information and facial landmarks along with default (color-blind) face detection algorithm. The skin regions are mapped using an artificial neural network (ANN). On public data sets, our face detector was more accurate than the available state-of-the-art engines, both commercial

Face matching utilized a SIPP approach using weighted image descriptor ensemble to optimize the matching accuracy without training. Our MEADOW key-point filtering, attribute bucketing and FLANN indexing helped speed-up queries up to 20-times (compared to the linear search), keeping turn-around time within one second for a typical real-world collection.

We have annotated thousands of face images in the PL dataset with face, profile and landmark regions. The annotated datasets are public domain and can be made available upon request.

We are currently researching the *human-in-the-loop* (HiL) approach for naturally merging face image retrieval with annotation, making both more efficient via semi-supervised and incremental machine learning techniques as well as via more natural human-computer interactions, which may include the development of more convenient game-like visual annotation tools, and use of crowd-sourcing for developing more comprehensive testing and evaluations data sets, including video, because mobile technology tends to generate an increasing amount of moving pictures often with characteristic audio tracks, quite useful for practical face and object image retrieval.

Our public FaceMatch services can be expanded in several ways including visual search by photo for missing children, pets, as well as detecting disaster scenes. FaceMatch R&D team is actively engaged in research and development that may lead to (i) robust automatic estimation of gender, age and ethnicity, and (ii) robust image retrieval depicting objects and animals.

REFERENCES

Ahonen, T., Hadid, A., & Pietikäinen, M. (2004). Face recognition with local binary patterns. In *Proceedings of the European Conference on Computer Vision* (pp. 469–481). Springer.

Azeem, A., Sharif, M., Raza, M., & Murtaza, M. (2014). A survey: face recognition techniques under partial occlusion. *Int. Arab J. Inf. Technol.*, *11*(1), 1–10.

Bay, H., Tuytelaars, T., & Gool, L. V. (2006). SURF: Speeded up robust features. In *European Conference on Computer Vision* (pp. 404–417).

Beham, M. P., & Roomi, S. M. M. (2013). A Review Of Face Recognition Methods. *International Journal of Pattern Recognition and Artificial Intelligence*, *27*(04), 1356005. https://doi.org/10.1142/S0218001413560053

Beveridge, J. R., Phillips, P. J., Bolme, D. S., Draper, B. ., Givens, G. H., Lui, Y. M., … Cheng, S. (2013). The challenge of face recognition from digital point-and-shoot cameras. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on* (pp. 1–8). https://doi.org/10.1109/BTAS.2013.6712704

Calonder, M., Lepetit, V., Strecha, C., & Fua, P. (2010). BRIEF: Binary Robust Independent Elementary Features. In *Proceedings of the 11th European Conference on Computer Vision: Part IV* (pp. 778–792). Berlin, Heidelberg: Springer-Verlag. Retrieved from http://dl.acm.org/citation.cfm?id=1888089.1888148

Candemir, S., Borovikov, E., Santosh, K. C., Antani, S. K., & Thoma, G. R. (2015). RSILC: Rotation- and Scale-Invariant, Line-based Color-aware descriptor. *Image Vision Computing*, *42*, 1–12. https://doi.org/10.1016/j.imavis.2015.06.010

Chum, O., Pajdla, T., & Sturm, P. (2005). The geometric error for homographies. *Computer Vision and Image Understanding*, *97*(1), 86–102. https://doi.org/http://dx.doi.org/10.1016/j.cviu.2004.03.004

Deng, P., & Pei, M. (2008). Multi-View Face Detection Based on AdaBoost and Skin Color. In *Intelligent Networks and Intelligent Systems, 2008. ICINIS ’08. First International Conference on* (pp. 457–460).

Dharani, T., & Aroquiaraj, I. L. (2013). A survey on content based image retrieval. In *Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on* (pp. 485–490). https://doi.org/10.1109/ICPRIME.2013.6496719

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/http://dx.doi.org/10.1016/j.patrec.2005.10.010

Gao, Y., & Qi, Y. (2005). Robust Visual Similarity Retrieval in Single Model Face Databases. *Pattern Recogn.*, *38*(7), 1009–1020. https://doi.org/10.1016/j.patcog.2004.12.006

Hsu, R.-L., Abdel-Mottaleb, M., & Jain, A. K. (2002). Face detection in color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 696–706.

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. University of Massachusetts, Amherst.

Jacobs, C. E., Finkelstein, A., & Salesin, D. H. (1995). Fast multiresolution image querying. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques* (pp. 277–286). New York, NY, USA: ACM.

Jafri, R., & Arabnia, H. R. (2009). A Survey of Face Recognition Techniques. *JiPS*, *5*(2), 41–68.

Jain, A. K., Klare, B., & Park, U. (2012). Face Matching and Retrieval in Forensics Applications. *MultiMedia, IEEE*, *19*(1), 20–20. https://doi.org/10.1109/MMUL.2012.4

Jain, V., & Learned-Miller, E. (2010). *FDDB: A Benchmark for Face Detection in Unconstrained Settings* (No. UM-CS-2010-009). University of Massachusetts, Amherst.

Jones, M., & Rehg, J. M. (2002). Statistical Color Models with Application to Skin Detection. In *International Journal of Computer Vision* (pp. 274–280). Retrieved from Michael.Jones@compaq.com

Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., & Brossard, E. (2016). The MegaFace Benchmark: 1 Million Faces for Recognition at Scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Kim, E., Huang, X., & Heflin, J. (2011). Finding VIPs - A visual image persons search using a content property reasoner and web ontology. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on* (pp. 1–7).

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Lu, C., & Tang, X. (2014). Surpassing Human-Level Face Verification Performance on LFW with GaussianFace. *CoRR*, *abs/1404.3840*.

Lui, Y. M., Bolme, D., Phillips, P. J., Beveridge, J. R., & Draper, B. A. (2012). Preliminary studies on the Good, the Bad, and the Ugly face recognition challenge problem. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 9–16). https://doi.org/10.1109/CVPRW.2012.6239209

Muja, M., & Lowe, D. G. (2009). Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Applications* (pp. 331–340).

Naruniec, J. (2010). A Survey on Facial Features Detection. *International Journal of Electronics and Telecommunications*, *56*(3), 267–272.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313.

Phillips, P. J., Beveridge, J. R., Draper, B. A., Givens, G., O'Toole, A. J., Bolme, D. S., … Weimer, S. (2011). An introduction to the good, the bad, & the ugly face recognition challenge problem. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* (pp. 346–353). IEEE.

Rosten, E., Porter, R., & Drummond, T. (2010). Faster and Better: A Machine Learning Approach to Corner Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 105–119. https://doi.org/10.1109/TPAMI.2008.275

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. In *IEEE International Conference on Computer Vision* (pp. 2564–2571).

Scheirer, W. J., Kumar, N., Iyer, V. N., Belhumeur, P. N., & Boult, T. E. (2013). How reliable are your visual attributes? In *Proceedits of SPIE* (Vol. 8712, p. 87120Q–87120Q–12). https://doi.org/10.1117/12.2018974

Sharif, M., Mohsin, S., & Javed, M. Y. (2012). A Survey: Face Recognition Techniques. *Research Journal of Applied Sciences, Engineering and Technology*, *4*(23), 4979–4990.

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). 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*.

Tan, X., Chen, S., Zhou, Z.-H., & Zhang, F. (2006). Face recognition from a single image per person: A survey. *Pattern Recognition*, *39*(9), 1725–1745.

Thoma, G., Antani, S., Gill, M., Pearson, G., & Neve, L. (2012). People Locator: a system for family reunification. *IT Professional*, *14*, 13–21.

Viola, P., & Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, *57*, 137–154.

Wolf, L., Hassner, T., & Taigman, Y. (2009). The one-shot similarity kernel. In *In International Conference on Computer Vision* (pp. 897–902). Retrieved from http://www.openu.ac.il/home/hassner/projects/Ossk

Zhang, C., & Zhang, Z. (2010). *A Survey of Recent Advances in Face detection*. Microsoft.

Zhou, H., Mian, A., Wei, L., Creighton, D., Hossny, M., & Nahavandi, S. (2014). Recent Advances on Singlemodal and Multimodal Face Recognition: A Survey. *Human-Machine Systems, IEEE Transactions on*, *44*(6), 701–716. https://doi.org/10.1109/THMS.2014.2340578

Zhu, X., & Ramanan, D. (2012). Face Detection, Pose Estimation, and Landmark Localization in the Wild. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Zuliani, M., Kenney, C. S., & Manjunath, B. S. (2005). The multiRANSAC algorithm and its application to detect planar homographies. In *IEEE International Conference on Image Processing*. Genova, Italy.