Human Saliency-Driven Patch-based Matching for Interpretable Post-mortem Iris Recognition

Aidan Boyd Daniel Moreira Andrey Kuehlkamp Kevin Bowyer Adam Czajka
University of Notre Dame, Notre Dame, IN, USA
{aboyd3, dhenriq1, akuehlka, kwb, aczajka}@nd.edu

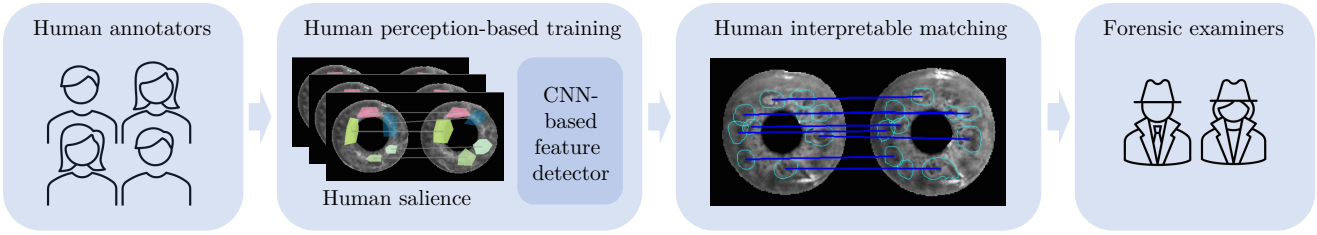


Figure 1: Our proposed human-interpretable forensic iris recognition benefits from using a saliency-driven feature detector trained with regions annotated by examiners solving previous forensic iris comparison tasks. Automatically-detected patches are compared and both the overall comparison score, as well as matching feature pairs, are presented to a human examiner.

Abstract

Forensic iris recognition, as opposed to live iris recognition, is an emerging research area that leverages the discriminative power of iris biometrics to aid human examiners in their efforts to identify deceased persons. As a machine learning-based technique in a predominantly human-controlled task, forensic recognition serves as "back-up" to human expertise in the task of post-mortem identification. As such, the machine learning model must be (a) interpretable, and (b) post-mortem-specific, to account for changes in decaying eye tissue. In this work, we propose a method that satisfies both requirements, and that approaches the creation of a post-mortem-specific feature extractor in a novel way employing human perception. We first train a deep learning-based feature detector on post-mortem iris images, using annotations of image regions highlighted by humans as salient for their decision making. In effect, the method learns interpretable features directly from humans, rather than purely data-driven features. Second, regional iris codes (again, with human-driven filtering kernels) are used to pair detected iris patches, which are translated into pairwise, patch-based comparison scores. In this way, our method presents human examiners with human-understandable visual cues in order to justify the identification decision and corresponding confidence score. When tested on a dataset of post-mortem iris images collected from 259 deceased subjects, the proposed method places among the three best iris comparison tools, demonstrating better results than the commercial (non-human-interpretable) VeriEye approach. We propose a unique post-mortem iris recognition method trained with human saliency to give fully-interpretable comparison outcomes for use in the context of forensic examination, achieving state-of-the-art recognition performance.

1. Introduction

The high entropy of iris texture patterns has allowed this modality to rank among the most reliable means of biometric identification of living individuals. Large-scale iris biometric applications include the national person identification and border security program Aadhaar in India (with over 1.2 billion pairs of irises enrolled) [17], the Homeland Advanced Recognition Technology (HART) in the US (up to 500 million identities) [13], and the NEXUS system, designed to speed up border crossings for pre-approved travelers between Canada and the US [8].

It was recently demonstrated that comparison of live irises with their post-mortem counterparts is feasible [31, 36, 37, 5], and – depending on environmental factors – can be viable even several weeks after death. This discovery opened new research areas in forensic iris recognition, with applications that can have huge impact in the context of

accidents, natural disasters and combat zones. While automatic image processing and comparison tools frequently support forensic examiners, the final decision is made by a human expert. This is why human interpretability of the algorithm's decisions is essential.

This paper proposes a human-interpretable iris recognition method designed specifically for post-mortem (forensic) iris recognition. The core novel component is the feature detection model trained using image regions annotated by human examiners as salient to their decision-making. Instead of narrowing down the interpretable feature selections to several known anatomical iris features (such as collarette, Fuchs' crypts, contraction folds, or Kruckman-Wolfflin bodies), we asked 283 humans to compare iris image pairs and to annotate *any* visual features within the iris annulus that support their decision on whether the image pair is a match. The deep learning-model was trained with human-annotated patches to locate iris image regions similar to those selected by humans. Those automatically-detected local regions are then compared using a modified (regional) version of the human-driven binary statistical image features descriptor [10]. Figure 1 illustrates the proposed method's pipeline.

This approach may seem to be similar to a traditional keypoint-based iris comparison, *e.g*. employing SIFT or SURF descriptors [3, 19], but the **fundamental difference from keypoint-based approaches** lies in the human-driven feature detector, which has several advantages. First, it detects features that are closer to those which human experts would choose. Since these features are more task-driven, they may have better discrimination power and thus a lower number of features is necessary for a high-confidence comparison (on the order of a dozen) compared to a general-purpose keypoint-based solutions (which needs hundreds of features to offer high-confidence comparison, as shown through experimentation). This makes the visualization less cluttered and more human-interpretable. Second, the proposed method detects regions that have specific shapes, and not just their central locations, which also aids human interpretability. Last but not least, the comparison performance of the proposed method obtained on a sequestered dataset of post-mortem iris images compares very favorably with state-of-the-art methods: it is slightly worse than one of the non-human-interpretable methods [10], but beats the commercial VeriEye comparison tool [28] as well as all human-interpretable iris recognition methods known to us, and included in these evaluations. In summary, the **main contributions** of this work are:

(a) A novel human-driven, human-interpretable iris regional-based comparison method, designed specifically for forensic applications;

(b) A database of human-annotated iris features, along with the human classification decisions for comparing pairs of post-mortem iris samples; in addition to this paper's reproducibility purposes, this data can serve as a useful resource for studying human-machine pairing in the context of forensic iris recognition;

(c) Trained models and source codes of the proposed method, able to be applied in both forensic and live human-interpretable iris recognition.

All resources (source codes, models, human annotations) are available at `https://github.com/CVRL/PBM`.

2. Related Work

Forensic Iris Recognition. Forensic iris recognition was long believed to be impossible, due to incorrect assumptions about the pupil dilation after death, the cornea becoming cloudy [11], or even the entire iris decaying only minutes after death [35]. These assumptions were debunked by Sansola [31], who demonstrated that perimortem (image acquired just before death) to postmortem iris comparison is possible, and who observed correct comparison results for at least 70% of cases when only postmortem images were compared (depending on time after death). Other groups confirmed the feasibility of forensic iris recognition, with time period after demise ranging from a few days [5] (outdoor conditions during summer) to several weeks [37, 32] (mortuary or winter-time outdoor conditions). Several post-mortem iris recognition datasets are available to researchers, created by Trokielewicz *et al*. [36, 37]. These datasets are accompanied by emerging, but non-human-interpretable, post-mortem iris recognition methods, following the well-known iris code approach (with domain-specific filtering kernels) [39, 38]. The existence of a survey devoted to post-mortem iris recognition [7] suggests that this research area has gained momentum, and results may contribute to large-scale forensic applications such as the FBI's Next Generation Identification (NGI) service [14].

Explainable and Region-Based Approaches. Iris recognition results have historically been opaque to human interpretability. Daugman's [12] mathematically-elegant explanation of iris recognition's high discriminative power does not lead to intepretability in the context of intuitive features easily recognized by humans. This has led to work on interpretable approaches that point a human examiner to elements of the comparison results that aid in explaining their decision of whether two images are from the same iris.

Active application of deep learning methods to iris recognition changed this situation rather marginally. Among various Convolutional Neural Network-based approaches known to us [22, 16, 25, 29, 43, 2, 41, 42, 6], only one proposed a human-interpretable output. This is in a form of Class Activation Map overlaid on post-mortem iris images to suggest to the human examiner regions that

were salient to the model [21]. There have been previous attempts that approach iris recognition via keypoint-based comparison, which can be more interpretable than iris codes. Important works include using Scale-Invariant Feature Transform (SIFT) for iris image retrieval [34], combining Speeded-Up Robust Features (SURF) with wavelet-based texture descriptors [19], and leveraging anatomical properties of the iris, such as crypts and anti-crypts [33]. None of the previous methods, however, were designed specifically for forensic applications.

Our work is different from past efforts in the following respects. One, our technique is designed specifically for post-mortem iris recognition, using a large dataset of images acquired from 430 deceased subjects. Two, our feature detector is trained using features annotated by humans as salient for their decision on comparison, and so it is guided toward detecting features salient for humans. Human-interpretable results are essential if iris comparison is to serve in forensic applications.

3. Features for Human Iris-Match Decisions

To understand what features are useful to human examiners for post-mortem iris comparison, we collected annotations from humans performing an iris-comparison task[1]. Following a process similar to the ACE-V protocol used in fingerprint comparison [40], data acquisition took place in two steps. The first step is *Match Evaluation*, as in the "evaluation" step in ACE-V, during which "the final determination as to whether a finding of individualization, or same source of origin, can be made" (cf. Sec. 9.3.2 and 9.3.3 in [40]). In this step, the collection of annotation data was made from subjects comparing iris images in the absence of any prior knowledge about the source of samples. The second step is *Match Verification*, as in the ACE-V "verification" stage, which is "independent examination by another qualified examiner resulting in the same conclusion" (cf. Sec. 9.3.5 in [40]). In this step, the annotations collected in the first step were presented to new subjects for them to either agree or disagree with, and to supply annotations supporting their decisions. The annotation tool for both step 1 and step 2 is shown in the supplementary materials.

Iris images used in our experiments are a combination of a publicly-available post-mortem iris dataset [37], and two datasets collected in a medical examiner's office, of which one has been submitted to the National Archive of Criminal Justice Data (NACJD) archives [1] and can be requested from the NACJD for research purposes. All the iris images used in evaluating our work are available to other researchers.

[1] All data collection was done under an approved IRB protocol that allows for distribution of the data to the research community.

Step 1: Match Evaluation. Subjects were presented with a pair of post-mortem iris images, and asked to decide whether the images are from the same eye or different eyes. Once this decision was made, they were asked to annotate features salient to their decision. If the decision was that the images are from the same eye, they were asked to annotate at least 5 pairs of corresponding features between the images. These will be referred to as matching features. If the decision was that the images are from different eyes, they were asked to highlight at least 5 regions on either iris that are present in one image but not the other. These are called non-matching features. There is also a "Don't Know" option to address inconclusive cases. When this option was selected, they could then annotate either non-matching features, matching features or combinations of both.

The step 1 data was collected from 152 human annotators. All of the annotators are individuals associated with the University of Notre Dame. Each annotator was presented with 20 image pairs: 10 from the same eye ("genuine" pairs) and 10 from different eyes ("impostor" pairs). The pairs were presented in an order randomized for each annotator. Within the genuine/impostor categories, sampling was performed based on the post-mortem interval (PMI, time in hours since death). Pairs were curated such that at least one of the eyes in the pair is from a low-PMI range to make sure that the matching of artefacts such as specular highlights and wrinkles is minimized as these appear less frequently in lower PMI samples. It is also a more likely scenario in practice that lower PMI images are compared to higher PMI images.

Step 2: Match Verification. In a verification trial, a new annotator is presented with an image pair from a previous (matching) trial, the previous annotator's decision for that image pair, and a random subset of the previous annotations. Some data cleaning was performed prior to the verification trials, to remove incorrectly-completed annotated pairs from the previous matching trial. These new annotators were asked to make the same decision as in the matching trial: do the two images come from the same eye, or eyes of different persons? In the same manner as a matching trial, annotators were also required to annotate five feature-match pairs/non-matching features. The annotator on a verification trial could agree or disagree with the results from the previous matching trial. The inclusion of the annotations from the previous matching trial should serve to highlight regions that lead to different decisions.

Note that an annotator participating in a sequence of verification trials sees results of previous match trials by different annotators. One research question behind the verification trials is to find out if knowing the results of a previous match trial for a particular pair of images leads to better annotations and more accurate classification.

The step 2 annotations were collected from 131 new subjects using Mechanical Turk (MTurk). Restrictions on the MTurk workers include that the worker (a) be an MTurk "Master", meaning they had an exceptional approval rating, and (b) be located in a native-English-speaking country, to reduce communication errors in the instructions or instructional video. The annotations were visually inspected to remove blatantly erroneous samples, resulting from: tool malfunction, obvious misunderstanding of the task, and apparent "speed-runners", who gave minimal effort to move through the task as quickly as possible. Of the 2620 pairs shown in the second round, 89 (just over 3%) were deemed unusable. For each matching-trial annotation, there is an acceptable-quality verification-trial annotation.

Table 1: Accuracy of human subjects comparing and annotating pairs of iris images in two steps of the experiment.

	Step 1 (evaluation)	**Step 2** (verification)
Overall	57.3%	60.9%
Genuine pairs	36.3%	34.9%
Impostor pairs	78.4%	86.9%
Inconclusive	8.9%	4.9%
Number of annotators	152	131

Annotation Results. As shown in Table 1, accuracy of the verification trials is higher than that of the matching trials, primarily due to impostor pairs of images being classified with higher accuracy in the verification step. Interestingly, for genuine pairs the accuracy decreased slightly from the step 1 (evaluation) trials to the step 2 (verification) trials. Also, the number of inconclusive decisions decreased greatly in the verification trials, from 8.9% to 4.9%.

The inclusion of the decision and annotations from a previous matching trial allowed annotators in a verification trial to make better-informed decisions. Annotators could either agree or disagree with the previous annotation, but the additional context increased the overall accuracy. Because invalid matching-trial annotations were removed, the annotations shown to verification-trial annotators were good examples of correct experimental procedure. Thus, with this guidance the quality of annotations in verification trials increased, as well as the overall accuracy.

4. Methodology

As the PMI of post-mortem samples increases, the iris surface area usable for recognition diminishes. This a result of the emergence of decomposition artifacts such as cloudiness or wrinkles [7]. As well, the circularity of the iris boundaries becomes compromised. Thus, to improve the robustness of post-mortem iris comparison, the proposed method does not use the traditional assumption of a circular (or elliptical) iris boundary, or that all the iris that is not-occluded by eyelids contains usable texture. These are factors that make post-mortem iris comparison different from and harder than traditional iris recognition.

One of our approaches to circumvent post-mortem deformations is – in addition to forensic iris-specific image segmentation – to try to detect features that are unaffected by the decomposition process. Our method detects small regions of usable iris texture in an image, similarly to what humans would do, and then represents these feature patches as unique feature descriptors. The set of feature descriptions for two iris images is then used for comparison.

The proposed solution consists of three components: the feature detector, the feature descriptor and the comparison scheme. The goal of the feature detector is to find usable iris texture regions as explained above. The goal of the feature descriptor is to represent the detected regions of iris texture such that they are easily discernible from each other. Given the set of feature descriptions for two iris images, the comparison scheme outputs a score for the degree of similarity of the irises. This proposed method is thus referred to as Patch-Based Matching (PBM).

The over-arching design goal for our method is to be visually understandable to human examiners. The network we use for feature detection returns a visual representation of where the features are located, and then both the feature description and comparison scheme together can show which features are being matched together. At the potential tradeoff of slightly lower accuracy, our approach is completely transparent about the regions of the images that support the match / non-match decision and presents these results in a human-interpretable way to examiners.

4.1. Databases

The first publicly-available dataset for post-mortem iris recognition is the Warsaw BioBase Postmortem Iris v2.0 (Warsaw v2.0) [37]. It consists of 1,200 near-infrared (NIR) post-mortem images from 37 unique cadavers in a mortuary environment. The PMI ranges from 5 to 800 hours.

Two additional datasets were used in this work, both acquired in an operational medical examiner's setting. The first of these, *DCMEO1*, collected by the Dutchess County Medical Examiner's Office (DCMEO), NY, contains 621 NIR images from 134 cadavers (254 distinct irises). Images were acquired in sessions of varying PMI, up to a maximum of 9 sessions and 284 hours after death. The second dataset, *DCMEO2*, collected also by the DCMEO, consists of $5,770$ NIR images from 259 subjects. The longest PMI in this dataset is 1,674 hours (69 days), captured at 53 different PMI sessions. Warsaw v2.0, *DCMEO1* and *DCMEO2*

are entirely subject-disjoint. Warsaw v2.0 is combined with *DCMEO1* to create what is referred to as the "combined dataset" used for training and validation, and to collect human decisions and annotations, as described in Sec. 3. *DCMEO2* is held out during training and validation and acts as a subject-disjoint test set. While the DCMEO1 is not publicly availble, the DCMEO2 set will be available at the NACJD [1].

4.2. Data Preprocessing

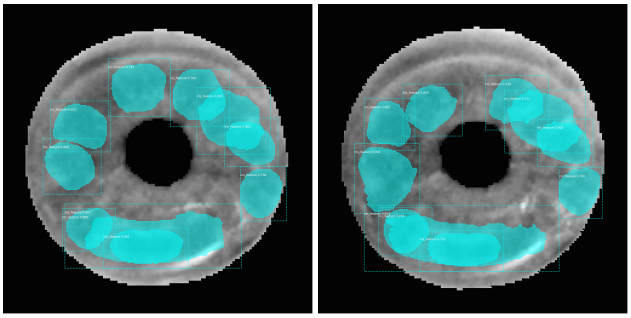
Annotation Data Preprocessing. For this training, only annotations from **correct decisions** about **genuine pairs** are shown to the network. Because each image was annotated multiple times in different pairs, bountiful human-derived ground truth is collected. It was decided that supplying the same image to the network with different annotations from each annotator that saw the image could hinder effective learning. Conversely, simply using all feature annotations for an image would be too much redundant information, as many people may have annotated the same feature. Thus, to conserve resources for training and remove redundant annotations, a method of aggregating to one ground truth set of annotations was applied. This was achieved by first collecting all sets of annotations for a given image. Next, we take all overlapping annotations and if there is an overlap of greater than 50% area, we remove the smaller feature annotation. This leaves us with the minimum feature set where the overlap between any two annotations is no greater than 50%. The resulting set of iris images with associated correct annotations contains 716 images, split in a subject-disjoint manner to end up with 518 images in the train set and 198 images in the validation set (70%/30% proportion).

Image Preprocessing. Using a post-mortem-iris-specific application of SegNet [39], the iris images are first segmented and cropped to 256×256 pixels around the detected iris. The segmentation mask is used to set all regions not corresponding to the iris texture to zero (black in the image). Contrast-limited adaptive histogram equalization (CLAHE) is applied to the cropped image to accentuate the iris texture, as reported in [27] to be an effective image enhancement means in case of forensic iris recognition.

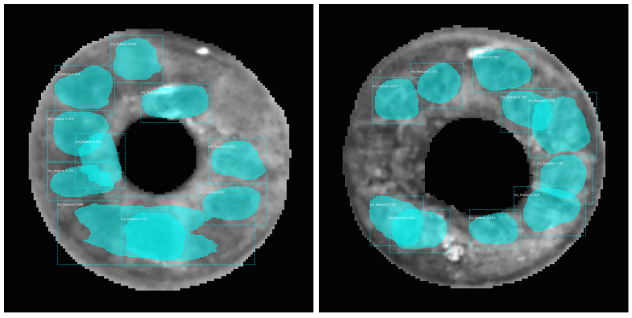
4.3. Feature Detection

The MaskRCNN instance segmentation architecture with a ResNet50 [18] backbone is trained to detect individual features present in the iris. In addition, a confidence score is also returned and can be used to rank the detected features. Two examples of iris images with detected features can be seen in Fig. 2. For each image, a maximum of 10 individual features are detected.

A noteworthy point is that the data is annotated in a pair setup, whereas only individual images are used for the MaskRCNN model training. The rationale is that if pairs are annotated rather than individual images, only features that can be used for comparison will be annotated. If all features were annotated, some might not be useful for recognition. The goal from this network is to determine regions with all good features for comparison.



(a) Genuine Iris Pair



(b) Impostor Iris Pair

Figure 2: Output of the human-guided MaskRCNN-based feature detector for a genuine pair (a) and an impostor pair (b). Cyan patches are automatically-detected features.

Experimental Parameters. Due to the limited size of the dataset, extensive augmentation is performed. The augmentations used include left-right flip, up-down flip, ± 30 degrees rotation and Gaussian blurring. This set of augmentations makes sense in the post-mortem iris recognition domain (for instance, up-down flips or severe rotations may happen when a deceased person is approached by an operator from different angles; this is almost not observed in case of live iris recognition, where subjects' eyes are usually positioned horizontally and aligned with the sensor). All layers in the model were trained for 10 epochs using a learning rate of 0.001. After 10 epochs, the learning rate is divided by 10 and the network head layers are fine-tuned for a further 10 epochs. The optimizer for the network was Stochastic Gradient Descent (SGD). To enable experimental replication, the MaskRCNN specific parameters used to train the model are included in the supplementary materials. The trained model weights are also released with this work.

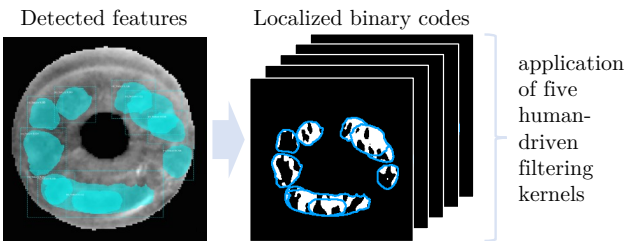


Figure 3: Visualizations of local iris patches encoded with a modified human-driven BSIF method.

4.4. Feature Description

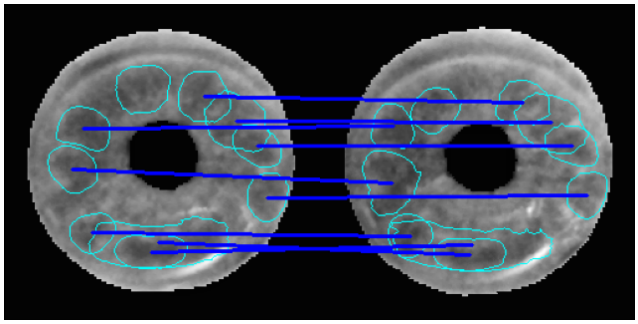
Based on the performance of post-mortem comparison tools in the recent survey by Boyd *et al*. [7], the best current performing method is using human-driven BSIF [10]. However, this approach performs iris comparison in a non-interpretable manner such that it is unclear what features present in both irises lead to a match. The proposed feature description for our work aims to leverage the proven performance of the human-driven BSIF method and combine it with our human-interpretable feature detection. In a traditional deep-learning based approach, input images need to be of a specific size, so either features must be extracted of that size only or resized. This ignores both the shape and scale of the feature and can lead to false matching. As the human-driven BSIF approach is not deep learning based, there is no size constraint on features and thus structural integrity of detected features is preserved.

To achieve this integration to our method, the cropped images are encoded in the same BSIF format as was found to be optimal by [10]. That is, we apply five filtering kernels of size 17×17 pixels, learned using eye-tracking data. Using the feature set detected with the MaskRCNN model, the feature description is the extracted region of that feature on the BSIF encoded image, as shown in Fig. 3.

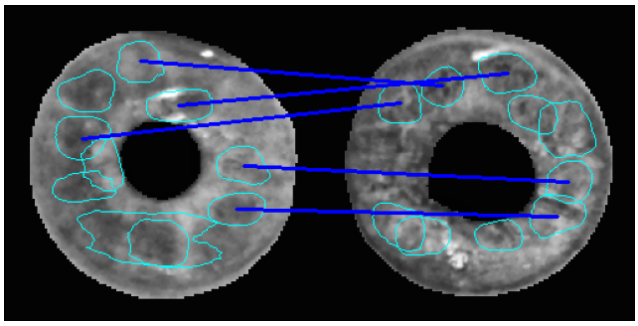
4.5. Matching Scheme

Given two sets of detected features, the first step is to determine the set of all valid matches. For a pair of features (one on each image) to be considered, they must be within ± 20 degrees of rotation of one another. This is determined by establishing the center point of both irises using the segmentation mask. Using the relative position of these iris centers to the center point of the detected features, the angle can be determined.

The distance metric used to determine the closeness of two features is the Hamming distance (HD) between the two feature descriptions (binary iris codes obtained for each filter). Because five filters are used to get the iris code, the final distance is the mean of the HDs calculated for each of the iris codes. For this distance calculation to work, features need to be the same size. Thus, the largest possible overlap between the two features is calculated and a full iteration of all possible combinations of overlapped features is performed. The smallest distance found in any iteration is accepted as the score for that pair. To insulate against edge cases, the area of the maximum possible overlap must be greater than 50% the area of the smaller feature.



(a) Genuine Iris Pair – Match



(b) Impostor Iris Pair – Non-match

Figure 4: Output of Patch-Based Matching. Pairs of features that are matched as being the most similar are linked by the dark blue lines. The genuine pair (a) shows parallel lines linking features resulting in a match, whereas the impostor pair (b) has crossing lines and a non-match result. The human examiner can quickly verify the algorithm's result by examining the proposed matching features.

Once the list of all valid matching features is established, it is reduced to ensure that each feature can only be used in one match. This is done by ranking all valid matches based on their increasing distance apart. Once a feature has been used in a pair, neither of those features can be used again in other pairings. The initial sorting ensures only the strongest pairings are maintained. The final comparison score for the two sets of detected features is the average distance of the five most similar feature pairs, or however many there are if less than five. Thus, the closer the score to zero, the more similar the feature pairs in the images and the more likely it is a genuine match. An example of the output of the tool can be seen in Fig. 4. Matching features are clearly articulated in a human-interpretable manner.

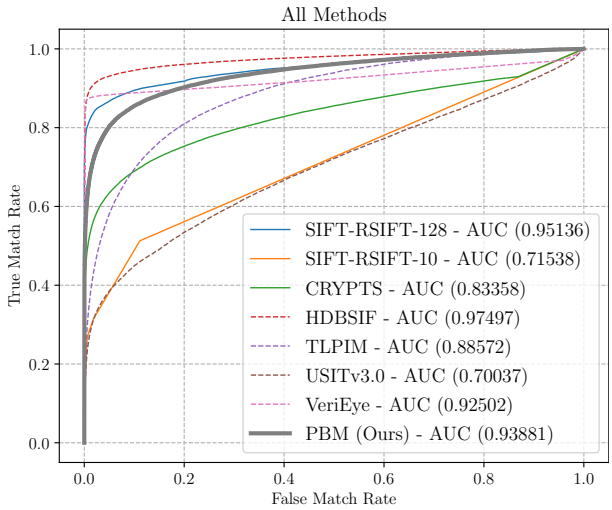


Figure 5: Results for all baselines plus our proposed method (PBM). Dashed lines represent non-interpretable methods, solid lines represent human-interpretable methods, thicker solid line is our PBM method.

5. Evaluation

Table 2: Equal Error Rates and decidability (d') scores. The proposed method compares favorably to the commercial (non-human-interpretable) VeriEye, and is best in terms of the d' score among all human-interpretable methods).

Method	EER (%)	d' score
SIFT-RSIFT-128	10.4	1.39
SIFT-RSIFT-10	35.4	0.98
Crypts [9]	23.2	0.94
HDBSIF [10]	**6.4**	**2.54**
TLPIM [21]	19.5	1.51
USIT v3.0 [30]	35.8	0.82
VeriEye [28]	11.1	1.29
PBM (proposed)	**12.8**	**2.08**

5.1. Algorithms Compared To

To compare the performance of the proposed method with state-of-the-art iris recognition, a set of baseline experiments were conducted with a variety of methods: human-interpretable and non-human-interpretable, deep learning and handcrafted, as well as commercial and open-source.

TLPIM (Triplet Loss Postmortem Iris Model) [21] is a deep-learning based post-mortem iris comparison approach that uses Class Activation Mapping to visualize important regions for post-mortem iris recognition.

VeriEye [28] is a popular commercial iris recognition tool produced by Neurotechnology. VeriEye uses Taylor expansion to extract image features that are then compared using a metric called "elastic similarity" in which impostor pairings produce results near zero.

USIT v3.0 [30] is an open-source academic tool that implements Daugman-style iris recognition, using iris codes to calculate the Hamming distance between images. The configuration as suggested by the USIT authors was used: segmentation using CAHT, feature extraction from Ma *et al*. [24] and comparison using Hamming distance.

HDBSIF (Human-Driven BSIF) [10] uses the ICA-trained filters, as in the original BSIF pipeline [20], for extracting iris features. Two core differences with an original BSIF pipeline are (a) filters trained on iris image patches extracted from an eye-tracking device for people comparing iris samples, and (b) using binarized filter responses directly as iris codes, instead of comparing histograms of BSIF codes.

Keypoint-based. SIFT-RSIFT is a combination of general-purpose SIFT [23] and a more accurate variation of its 128-dimensional feature descriptor, a.k.a. Root SIFT [4]. This solution leverages geometrically consistent content comparison [26] to find pairs of keypoints across two compared irises that present small feature-wise L_2 distances and high position equivalence (a.k.a. matches). Inspired by fingerprint minutiae comparison [15], the number of matches is used to express the similarity between the two irises. Genuine pairs are expected to present large numbers of matches, while impostors are expected to present small numbers. To obtain robustness to the post-mortem collapse, keypoints are extracted from normalized irises. In the experiments, we explore two numbers of extracted keypoints, 128 (SIFT-RSIFT-128) and ten (SIFT-RSIFT-10).

Crypts [9] method implements detection and automatic comparison of the iris crypts – features that can be easily interpreted by humans. The designed comparison scheme is able to handle potential topological changes in the detection of the same crypt in different images.

5.2. Results

From the ROC curves in Fig. 5, the best performing method on the held-out testing set (*DCMEO2*) is HDBSIF. This is consistent with the results in [7] for post-mortem iris recognition. However, HDBSIF is not human-interpretable, as it gives no indication of what features were important in the comparison. Instead, it uses the entire available iris surface for the comparison. Thus, it is not directly applicable to a forensic scenario in which human examiner's assessment is needed and the explanation as to why the pair of irises match, or not, is critically important.

The best performing method that supplies justification of the decision is SIFT-RSIFT-128. This method returns pairs of matching keypoints and singular non-matching keypoints from the two sets of 128 keypoints. The AUC achieved by

this method is 0.951, narrowly outperforming our approach that attains an AUC of 0.939. However, because SIFT-RSIFT-128 uses 128 keypoints on each iris, the comparison visualization may become cluttered and the interpretability is reduced due to the large number of extracted regions. Moreover, as in typical SIFT-like approaches, keypoints are represented by their central locations and compulsorily regular neighborhoods (usually circular or rectangular, thus ignoring the shapes of the compared regions during comparison), making them not anatomy-driven and less human-interpretable. The characteristics of being general-purpose and describing regular neighborhoods also reduce the robustness of keypoint-based approaches to the pupil dynamicity and post-mortem collapse of the irises (which are both visually non-linear and complex phenomena). As a consequence, keypoint-based solutions must be applied over normalized irises, again hindering human-interpretability. When the number of keypoints in the SIFT-RSIFT approach is reduced to be the same as for PBM, and thus its results are less cluttered, the performance decreases significantly to an AUC of 0.715 (see SIFT-RSIFT-10). Lastly, SIFT-RSIFT keypoints are not iris inspired and thus may not appear salient to a human examiner.

In our proposed PBM method, the feature extractor is trained from human-annotated iris patches used in comparison. In addition to generating more human-understandable features, this apparently brings a set of very discriminative features. Surprisingly, the PBM approach outperforms deep learning-based method (*TLPIM*), commercial (*VeriEye*) and Daugman-like approaches (*USITv3.0*), while also displaying interpretablity. The closest method in terms of interpretability to PBM is *Crypts*. However, as seen in Fig. 5 the performance of *Crypts* is significantly worse than PBM in the post-morterm iris recognition regime (note that the Crypts method was desined for live, not cadaver irises, hence its lower performance is understandable). Additionally, the top-performing baselines (*HDBSIF & SIFT-RSIFT*) use iris normalization, which reduces interpretability as the iris texture is transformed to polar coordinate system. Our patch-based comparison works with original images as if it were performed by a forensic examiner.

Tab. 2 shows the d-prime values for all methods. The d-prime metric measures the separability between the mean of the genuine comparisons and impostor comparisons, with a higher value indicating better separation and thus better performance. A higher d-prime also means more consistency across matches and more reliability. The best-performing method with regards to d-prime is again HDBSIF. The best-performing interpretable method and second best overall is the proposed Patch-Based Matching. This shows that although there is a sacrifice in performance compared to HDBSIF, the PBM approach adds human interpretability yet still performs reliably and predictably.

6. Conclusions

This work introduces a new algorithm for post-mortem iris recognition, designed to (1) produce human-interpretable results and (2) achieve high accuracy in post-mortem iris comparison. Our foundation for producing human-interpretable results is a two-stage experimental data collection in which human examiners decide if a pair of iris images is from the same eye or not, and annotate image regions that support their decision. We find that the decisions of the verification stage are more accurate than those of the initial matching stage, with the improvement coming from more accurate classification of impostor pairs. The decisions of the verification stage also have a much lower frequency of "unsure" results. This implies that some large fraction of "unsure" judgements are examiner-dependent rather than a general result of available image features.

This experiment produces important and useful conclusions on its own concerning how to achieve high accuracy in human evaluation of pairs of iris images. And the iris image annotations collected in this experiment enable the training of the deep CNN to detect iris image features that are natural to human interpretation. Our proposed algorithm for automated comparison of post-mortem iris images is evaluated against various state-of-the-art methods, some traditional, some designed for human interpretability, and some designed for post-mortem iris comparison. Comparing algorithms on a publicly-available dataset of post-mortem iris images, our proposed algorithm achieves the second-highest d-prime among the algorithms evaluated. However, the algorithm with the highest d-prime uses a much larger number of features, and the features are not as directly human-interpretable.

Achieving a useful level of human interpretability almost always involves some tradeoff with accuracy. In forensic post-mortem iris comparison, human interpretability is essential. Our proposed approach demonstrates minimal tradeoff with accuracy in the context of post-mortem iris recognition, while being designed from ground-up to display human-interpretable feature regions and comparison. Source codes of the proposed method and trained models are being made available with this paper to contribute to the biometric community with human-interpretable, forensic-specific open-source iris recognition methods.

Acknowledgements

This work was supported by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, under Award 2018-DU-BX-0215. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

References

[1] National Archive of Criminal Justice Data archives – The source for data on crime and justice. https://www.icpsr.umich.edu/web/pages/NACJD/index.html. Accessed: 2022-11-18.

[2] Sohaib Ahmad and Benjamin Fuller. Thirdeye: Triplet based iris recognition without normalization. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9, 2019.

[3] Fernando Alonso-Fernandez, Pedro Tome-Gonzalez, Virginia Ruiz-Albacete, and Javier Ortega-Garcia. Iris recognition based on SIFT features. In *2009 First IEEE International Conference on Biometrics, Identity and Security (BIdS)*, pages 1–8, 2009.

[4] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2911–2918, 2012.

[5] David S. Bolme, Ryan A. Tokola, Chris B. Boehnen, Tiffany B. Saul, Kelly A. Sauerwein, and Dawnie Wolfe Steadman. Impact of environmental factors on biometric matching during human decomposition. In *IEEE Int. Conf. on Biometrics: Theory Applications and Systems (BTAS)*, pages 1–8, Niagara Falls, NY, USA, Sept 2016. IEEE.

[6] Aidan Boyd, Adam Czajka, and Kevin Bowyer. Are gabor kernels optimal for iris recognition? In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9, 2020.

[7] Aidan Boyd, Shivangi Yadav, Thomas Swearingen, Andrey Kuehlkamp, Mateusz Trokielewicz, Eric Benjamin, Piotr Maciejewicz, Dennis Chute, Arun Ross, Patrick Flynn, Kevin Bowyer, and Adam Czajka. Post-mortem iris recognition—a survey and assessment of the state of the art. *IEEE Access*, 8:136570–136593, 2020.

[8] Canada Border Services Agency and U.S. Customs and Border Protection. NEXUS. https://www.cbsa-asfc.gc.ca/prog/nexus/menu-eng.html. July 2021, accessed on April 21, 2022.

[9] Jianxu Chen, Feng Shen, Danny Ziyi Chen, and Patrick J. Flynn. Iris recognition based on human-interpretable features. *IEEE Transactions on Information Forensics and Security*, 11(7):1476–1485, 2016.

[10] Adam Czajka, Daniel Moreira, Kevin Bowyer, and Patrick Flynn. Domain-specific human-inspired binarized statistical image features for iris recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 959–967, 2019.

[11] John Daugman. BBC News: The eyes have it. http://news.bbc.co.uk/2/hi/science/nature/1477655.stm. Accessed: 2021-08-14.

[12] John Daugman. Information theory and the iriscode. *IEEE Transactions on Information Forensics and Security*, 11(2):400–409, 2016.

[13] Department of Homeland Security. Homeland Advanced Recognition Technology (HART). https://www.dhs.gov/publication/dhsobimpia-004-homeland-advanced-recognition-technology-system-hart-increment-1, February 2020 (accessed April 21, 2022).

[14] Federal Bureau of Investigation. Next Generation Identification (NGI). https://www.fbi.gov/services/cjis/fingerprints-and-other-biometrics/ngi, accessed on April 23, 2022.

[15] Francis Galton. *Fingerprint directories*. Macmillan and Company, 1895.

[16] A. Gangwar and Akanksha Joshi. DeepIrisNet: Deep iris representation with applications in iris recognition and cross-sensor iris recognition. *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2301–2305, 2016.

[17] Government of India. Unique Identification Authority of India. https://uidai.gov.in/, accessed on April 21, 2022.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Anca Ignat and Ioan Păvăloi. Experiments on iris recognition using SURF descriptors, texture and a repetitive method. *Procedia Computer Science*, 176:175–184, 2020. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020.

[20] Juho Kannala and Esa Rahtu. Bsif: Binarized statistical image features. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1363–1366, 2012.

[21] Andrey Kuehlkamp, Aidan Boyd, Adam Czajka, Kevin Bowyer, Patrick Flynn, Dennis Chute, and Eric Benjamin. Interpretable deep learning-based forensic iris segmentation and recognition. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 359–368, 2022.

[22] N. Liu, Man-Lian Zhang, Haiqing Li, Zhenan Sun, and T. Tan. DeepIris: Learning pairwise filter bank for heterogeneous iris verification. *Pattern Recognition Letters*, 82:154–161, 2016.

[23] David Lowe. Distinctive image features from scale-invariant keypoints. *Springer International Journal of Computer Vision*, 60(2):91–110, 2004.

[24] Li Ma, Tieniu Tan, Yunhong Wang, and Dexin Zhang. Efficient iris recognition by characterizing key local variations. *IEEE Transactions on Image processing*, 13(6):739–750, 2004.

[25] Shervin Minaee, AmirAli Abdolrashidi, and Yao Wang. An experimental study of deep convolutional features for iris recognition. *2016 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–6, 2016.

[26] Daniel Moreira, Aparna Bharati, Joel Brogan, Allan Pinto, Michael Parowski, Kevin Bowyer, Patrick Flynn, Anderson Rocha, and Walter Scheirer. Image provenance analysis at scale. *IEEE Transactions on Image Processing*, 27(12):6109–6123, 2018.

[27] Daniel Moreira, Mateusz Trokielewicz, Adam Czajka, Kevin Bowyer, and Patrick Flynn. Performance of humans in iris recognition: The impact of iris condition and annotation-driven verification. In *2019 IEEE Winter Conference on*

Applications of Computer Vision (WACV), pages 941–949, 2019.

[28] Neurotechnology. VeriEye SDK. https://www.neurotechnology.com/verieye.html. Accessed: 2024-04-27.

[29] Kien Nguyen, Clinton Fookes, Arun Ross, and Sridha Sridharan. Iris recognition with off-the-shelf cnn features: A deep learning perspective. *IEEE Access*, 6:18848–18855, 2018.

[30] Christian Rathgeb, Andreas Uhl, Peter Wild, and Heinz Hofbauer. Design decisions for an iris recognition SDK. In *Handbook of iris recognition*, pages 359–396. Springer, 2016.

[31] Alora Sansola. Postmortem iris recognition and its application in human identification. Master's thesis, Boston University, Boston, MA, USA, 2015.

[32] Kelly Sauerwein, Tiffany B. Saul, Dawnie Wolfe Steadman, and Chris B. Boehnen. The effect of decomposition on the efficacy of biometrics for positive identification. *Journal of Forensic Sciences*, 62(6):1599–1602, 2017.

[33] Feng Shen and Patrick J. Flynn. Iris crypts: Multi-scale detection and shape-based matching. In *IEEE Winter Conference on Applications of Computer Vision*, pages 977–983, 2014.

[34] Manisha Sam Sunder and Arun Ross. Iris Image Retrieval Based on Macro-features. In *2010 20th International Conference on Pattern Recognition*, pages 1318–1321, 2010.

[35] Adam Szczepański, Krzysztof Misztal, and Khalid Saeed. Pupil and iris detection algorithm for near-infrared capture devices. In Khalid Saeed and Václav Snášel, editors, *Computer Information Systems and Industrial Management*, pages 141–150, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[36] Mateusz Trokielewicz, Adam Czajka, and Piotr Maciejewicz. Post-mortem human iris recognition. In *2016 International Conference on Biometrics (ICB)*, pages 1–6, 2016.

[37] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Iris recognition after death. *IEEE Transactions on Information Forensics and Security*, 14(6):1501–1514, June 2019.

[38] Mateusz Trokielewicz, Adam Czajka, and Piotr Maciejewicz. Post-mortem iris recognition resistant to biological eye decay processes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2307–2315, 2020.

[39] Mateusz Trokielewicz, Adam Czajka, and Piotr Maciejewicz. Post-mortem iris recognition with deep-learning-based image segmentation. *Image and Vision Computing*, 94:103866, 2020.

[40] John R. VanderKolk. Examination Process, Chapter 9 in: Fingerprint Sourcebook. *National Institute of Justic*, 2011.

[41] Xiao Wang, Hui Zhang, Jing Liu, Lihu Xiao, Zhaofeng He, Liang Liu, and Pengrui Duan. Iris Image Super Resolution Based on GANs with Adversarial Triplets. In *Chinese Conference on Biometric Recognition (CCBR)*, 2019.

[42] Kai Yang, Zihao Xu, and Jingjing Fei. DualSANet: Dual Spatial Attention Network for Iris Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 889–897, January 2021.

[43] Zijing Zhao and Ajay Kumar. Towards more accurate iris recognition using deeply learned spatially corresponding features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.