

Bridging Human Concepts and Computer Vision for Explainable Face Verification

Miriam Doh^{1,2,†}, Caroline Mazini Rodrigues^{3,4}, Nicolas Boutry³, Laurent Najman⁴, Matei Mancas¹ and Hugues Bersini²

¹ISIA lab - Université de Mons (UMONS), Bd Dolez 31, 7000, Mons, Belgium

²IRIDIA lab - Université Libre de Bruxelles (ULB), Av. Adolphe Buyl 87, 1060, Ixelles, Belgium

³LRE - Laboratoire de Recherche de l'EPITA, 14-16 Rue Voltaire, 94270, Kremlin-Bicêtre, France

⁴LIGM - Laboratoire d'Informatique Gaspard-Monge, Université Gustave-Eiffel, 77454, Marne-la-Vallée, France

Abstract

With Artificial Intelligence (AI) influencing the decision-making process of sensitive applications such as *Face Verification*, it is fundamental to ensure the transparency, fairness, and accountability of decisions. Although Explainable Artificial Intelligence (XAI) techniques exist to clarify AI decisions, it is equally important to provide interpretability of these decisions to humans. In this paper, we present an approach to combine computer and human vision to increase the explanation's interpretability of a face verification algorithm. In particular, we are inspired by the human perceptual process to understand how machines perceive face's human-semantic areas during face comparison tasks. We use Mediapipe, which provides a segmentation technique that identifies distinct human-semantic facial regions, enabling the machine's perception analysis. Additionally, we adapted two model-agnostic algorithms to provide human-interpretable insights into the decision-making processes.

Keywords

Face verification, Explainable AI (XAI), Interpretability

1. Introduction

Face verification [1] aims to confirm an individual's identity based on facial features, with applications in law enforcement [2], border control [3], or smartphone security [4]. As AI becomes prevalent in decision-making [5], ensuring model fairness, accountability, confidentiality, and transparency is crucial [6]. However, complex ML models are often seen as 'black boxes' [7]. Explainable AI (XAI) [8] addresses this challenge by enhancing AI interpretability to make AI systems transparent and understandable to humans, thereby increasing trust in their decisions.

Saliency maps have become the most popular XAI solution in computer vision, offering insights into the critical features considered in the decision-making process. However, in face verification, decisions often rely on adjustable thresholds based on the specific application rather than understandable semantic classes. This raises questions about the adequacy of identifying the most important features in an image as the only explanation [9]. Taking inspiration from the human perceptual process, we propose a model-agnostic approach capable of determining

BEWARE-23 Joint Workshop @ AIXIA 2023, Rome, Italy, November 6th, 2023, Rome, Italy

[†] work supported by the ARIAC project (No. 2010235), funded by the Service Public de Wallonie (SPW Recherche).



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

how the machine perceives similar semantic areas of the face when comparing two faces. Our primary objective is to translate the XAI solution into human decision-making meaningfully. However, incorporating human-based semantics in the models' explanation process can also introduce human bias to these same explanations. To increase human interpretability, we must also assure the *Faithfulness* of explanations to the model's reasoning. *Faithfulness* refers to whether a feature, considered important for the model, changes the model's decision [10].

For face verification, the model extracts features for each face that will be compared. Modifications in the features will also impact how similar are the two faces. Therefore, it is essential to understand how face parts, such as an eye, would impact the final features.

To translate the model's knowledge to human knowledge as smoothly as possible, we first perform the segmentation of face parts based on human semantics. By considering the impact of those face parts on a set of face images, we can have a global view of the model's knowledge. Following the features' extraction (through the model), we verify if two people are the same by comparing their facial features. To understand the contribution of the chosen *concepts* to the relation between two compared faces, we introduce an algorithm grounded in the perturbation of facial regions linked to the extracted concepts, mirroring the human perceptual process of face recognition. It encompasses evaluating corresponding semantic areas along a spectrum of similarities, providing interpretation and contextualisation.

We structured this paper as follows: in Section 2, we present state-of-the-art methods for explaining the face verification task; in Section 3, we describe our framework, including the model concept's extraction and the perturbation methods for face comparison; in Section 4 we include the experimental results and limitations; in Section 5 we conclude the work.

2. Related work

Saliency maps, such as CAMs [11, 12] and RISE [13], are crucial for interpreting deep-learning models, revealing their inner workings. However, their primary development has centered on object recognition, leaving the field of face analysis relatively unexplored.

Despite its critical applications, research in face analysis has been limited. Works by [14, 15, 16] mainly focus on individual pixel or low-level feature significance, which can be challenging for human analysts and may not align with intuition. Conversely, LIME [17] employs superpixels within the image, providing a user-friendly, concept-driven explanation. However, this technique relies on a new model approximating the original, potentially obscuring the actual reasons for the original model's behavior [18].

Alternative approaches, such as TCAV [19] and knowledge graphss [20], prioritize low-level importance from pixels and aim to represent the model's knowledge through *concepts*. TCAV employs semantic concepts defined by users or discovered through image segment activations (with method ACE [21]), while knowledge graphs identify repeating patterns across network layers. Additionally, Tan et al. [22] introduced the Locality Guided Neural Network (LGNN), designed to induce filter topology that enhances the visualization of concepts.

Inspired by these methods, our approach combines human and model perspectives to identify essential concepts for face verification. We acknowledge that relying solely on human concepts can introduce bias while relying solely on the model can complicate interpretation.

3. Proposed Method

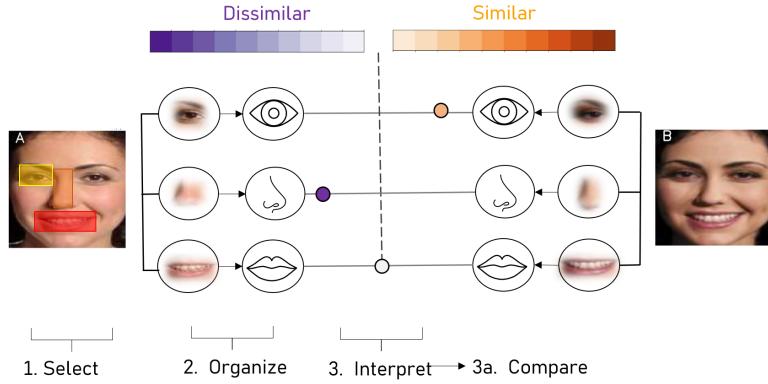


Figure 1: Face verification adaptation of XAI Perceptual processing framework proposed by [23] and inspired by how humans process stimuli (select, organize, interpret and compare)

To help humans understand how AI systems make decisions, it is essential to present the information in a way that aligns with human cognitive processes. Cognitive psychology provides valuable insights into how people perceive and process information when identifying faces. Taking inspiration from the flowchart proposed by [23], we aim to apply a similar method to face verification (see Figure 1). The human perceptual process consists of three key phases: selection, organization, and interpretation [24]. Cognitive psychology has shown that when recognizing faces, our attention is particularly drawn to particular facial areas, such as the eyes and nose [25, 26, 27]. Subsequently, in the perceptual process, these facial stimuli are organized into meaningful concepts, adding semantics to the process. Our brains compare these higher-level concepts to assess the similarity between items, facilitating face categorization. This comparative analysis may involve matching a face to a remembered image or with another face in front of us. In this context, we question the adequacy of salience maps used in computer vision as an explanation and their alignment with our human reasoning processes.

Based on cognitive psychology, we have developed a general flowchart shown in Figure 2. Generally, face verification systems rely primarily on a matching score between two face images A and B . This score, S_B^A , is computed using cosine similarity, which compares the feature vectors \mathbf{f}_A and \mathbf{f}_B extracted from each image as follows:

$$S_B^A = \frac{\mathbf{f}_A \cdot \mathbf{f}_B}{\|\mathbf{f}_A\| \|\mathbf{f}_B\|} \quad (1)$$

The resulting score ranges from 0 to 1, with a score of 1 indicating identical images ($A = B$). As our approach is model-agnostic, we aim to explain the algorithm by perturbing the inputs to study the system's decision behaviour concerning the input-output relationship. Inspired by the work of [16], our desired output is a similarity map indicating which face areas are considered similar or dissimilar for both images, using an AI model as a feature extractor. To achieve this, we perform semantic perturbation on images A and B , resulting in new images denoted as $A^{(n)}$ and $B^{(n)}$ where the n section is removed in both images. We obtain a new $S_{B^{(n)}}^{A^{(n)}}$ score from these

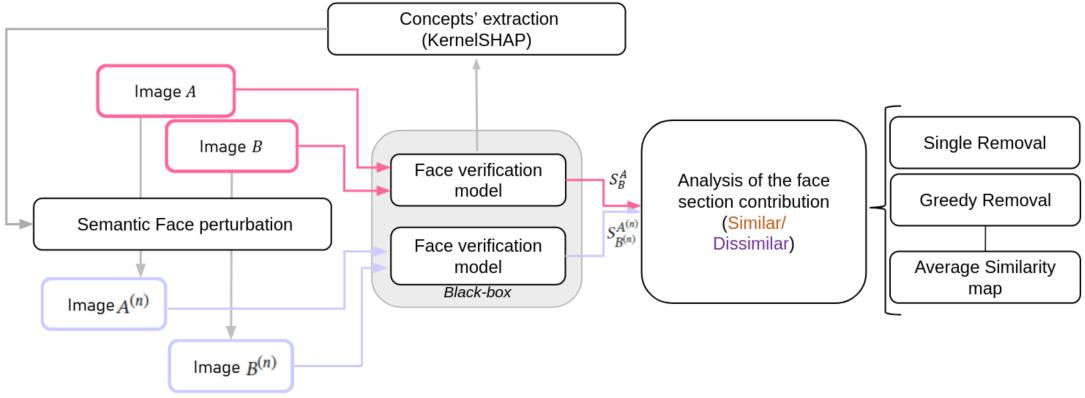


Figure 2: Proposed flowchart. We extract concepts from the face verification model (using KernelSHAP) and input them into a Semantic Face perturbation phase. In this phase, the two images' perturbation is made in the same regions to evaluate similarities and dissimilarities. We propose three algorithms for the perturbations: Single removal, greedy removal, and average similarity map.

perturbed images. By fairly masking the images, we can assess if the system perceives semantic areas, such as the eyes, as similar or dissimilar. Considering Δ_S the difference between original and new scores represented by Equation 2, if the $S_{B(n)}^{A(n)}$ decreases compared to S_B^A , it suggests that the removed parts positively contribute to the similarity ($\Delta_S \geq 0$). Conversely, its increase indicates a negative contribution ($\Delta_S < 0$).

$$\Delta_S = S_B^A - S_{B(n)}^{A(n)} \quad (2)$$

Compared to [16], our objective is to incorporate semantic masking in the perturbation process to increase interpretability by providing not only a map but also a chart related to the semantic areas. We apply two types of perturbation algorithms inspired by [15], allowing us to study the face section area's single or collaborative contributions and then incorporate this information within an average similarity map. This single/collaborative approach aligns with the notion that humans perceive and interpret faces in a relational/configurational way [28] (see Figure 3). First-order features concern individual components that can be processed independently (e.g., eyes, nose). Second-order features involve information acquired when simultaneously processing two or more parts together (e.g., spacing between eyes). Furthermore, higher-order features emerge from combinations of multiple first-order and/or second-order features. In our case, the single removal procedure models the information associated with first-order or single features, and the greedy removal procedure addresses the second-order features, wherein multiple parts are processed collectively.

3.1. Semantic Extraction

To incorporate semantics, we employ Mediapipe Face masks, a versatile open-source framework by Google, widely recognized for its face detection and landmark estimation capabilities. By ex-

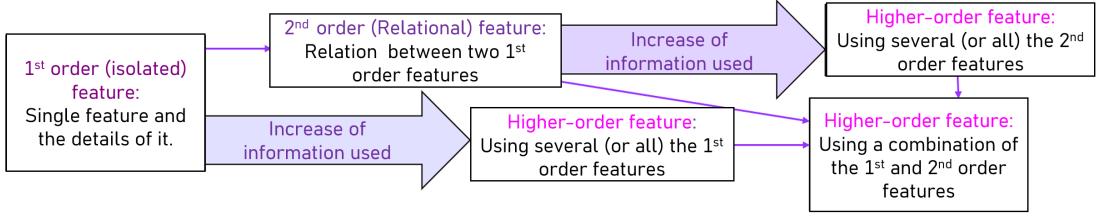


Figure 3: An interpretation of a relational/configural model of face perception.

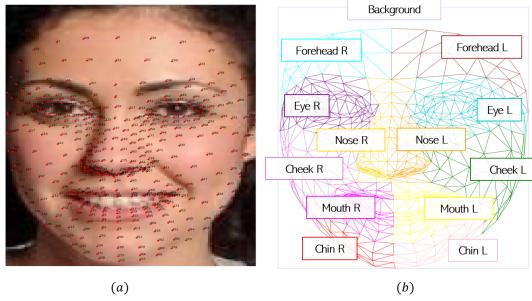


Figure 4: In the image (a) Mediapipe landmarks are plotted on the sample image. In the image (b), the 13 semantic sections are defined through the landmarks

tracting landmarks from Mediapipe, we defined 13 polygons corresponding to distinct semantic areas of the face (see Figure 4.a). The landmark estimation provided by Mediapipe is limited to specific facial regions, and hair or ears were not included in the earlier facial subdivisions. Nevertheless, this decision is consistent with previous research [29], which demonstrated that some areas of the face are more influential than others. For example, removing the ears has less impact on the final score than the eye area. Hence, we assumed these areas were not primarily influential and did not include them in our face classes. Additionally, face verification algorithms typically apply a preprocessing step for extracting the face area. Therefore, we reduce the area outside the face by applying MCTNN [30], a deep learning-based face detection algorithm. Overall, our subdivision of the face detected 13 distinct semantic classes, including the background (figure 4.b). With this approach, the semantic areas vary in size, resulting in larger maps having a more significant influence on the score than smaller ones. To mitigate this undesired effect, we introduce a weight, denoted as $w_{A,n}$ related to the section $n \in [1, m]$ with $m = 13$. The $w_{A,p}$ is defined as the rapport between the total area of the image A ($Area_A$) and the area of the mask $M^{(A,n)}$, $Area_{M^{(A,n)}}$, indicating region p (white pixels in the mask). This weight serves to counterbalance the differences in magnitude. Moreover, due to the precise face positioning achieved by Mediapipe, the masks obtained on images A and B may only partially match. This discrepancy arises because the depicted faces may not have the same position and expressiveness. For this reason, we define two weight $w_{A,n} = W(A, M^{(A,n)}) = \frac{Area_A}{Area_{M^{(A,n)}}}$ associated to A and $w_{B,n} = W(B, M^{(B,n)}) = \frac{Area_B}{Area_{M^{(B,n)}}}$ to B .

$$\hat{W}_{(A,B)_n} = \frac{w_{A,n} \cdot w_{B,n}}{\sum_{i=1}^m w_{A,i} \cdot w_{B,i}} \quad (3)$$

$$C_n = \Delta_S \cdot \hat{W}_{(A,B)_n} \quad (4)$$

In this manner, the contribution of the mask, defined as C_n , is weighted by $w_{A,n}$ and $w_{B,n}$, representing the relative weights associated with $M^{(A,n)}$ and $M^{(B,n)}$ masks, respectively.

3.1.1. Concepts Extraction

Using Mediapipe for face part extraction provides a human-based semantic segmentation, yet it may not align with how models perceive faces. To bridge this gap we introduce a model's concept extraction process. This involves filtering machine-important parts based on human semantics. For evaluating the importance of facial parts, we employ KernelSHAP [31], which combines LIME [17]'s interpretable components with Shapley values [32] from game theory which look for each feature contribution to the final result. We extract model importance scores for each of semantic parts. In face final representations with 512 features, for example, we will have 512 importance scores per human-semantics part. In the process of face verification, every feature change, negative or positive, is significant to determine faces' similarity, with emphasis on the magnitude of the change, instead of on the signal. If one feature of a human-semantic part obtained a negative Shap value, the lack of this part reduced the feature value, and vice versa. Therefore, negative and positive Shap values are equally important in our context. For this reason, sum the absolute Shap values throughout all the representation features to obtain a single importance value per part.



Figure 5: Examples of two images' human-semantics part importance scores using KernelSHAP [31]. We analyse two models: CasiaNet [33] in (a) and (c), and VGGfaces2 [34] in (b) and (d). Green parts are more important according to Shap scores. There are differences between important parts for different images, especially for VGGfaces2. That is why we aggregate ranked importance over 200 images.

Ultimately, we will have one importance value per semantic part (see Figure 5). However, this remains a local importance, i.e., an importance score according to a single image dataset. To increase globalism in the concepts' extraction, we need to include information from multiple images. Our solution is to combine the importance levels from a set of images by a ranking combination strategy. Each image obtains 13 importance scores (one per human-semantics part) that we can order. More significant scores are at the top of a ranking, as they were considered more important for the model. We use 200 images from CelebA [35] dataset to obtain 200 rankings. From these rankings' combinations, by a vote-based technique using BORDA count [36], we obtain a final ranking with more globally important concepts at the top.

The experiments will focus on the model's top eight concepts determined by this process.

3.1.2. Proposed similarity maps

The algorithm used to generate the similarity maps draws inspiration from the work of [15], where six algorithms were presented to create saliency maps. Specifically, we will employ the single removal approach (S0) and the greedy removal approach (S1), with the possibility of creating an average map of these two approaches (S_{AVG}). Our approach incorporates significant changes compared to previous research. First and foremost, we utilize semantically meaningful masks to perturb the images, diverging from conventional circular or square masks with a fixed shape. Moreover, since our objective is to generate community similarity maps between the two images, both images undergo perturbation, contrary to previous approaches that typically perturb only one of the images, thus aligning more closely with the strategy proposed by [16].

3.1.3. Single Removal - S0

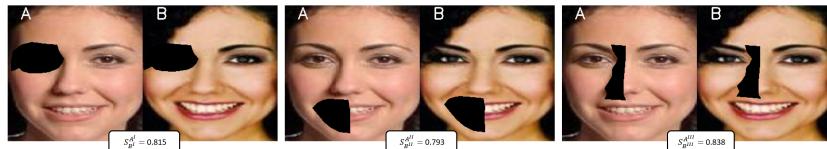


Figure 6: Samples of single removal where $S_{B^{(n)}}^{A(n)}$ is the cosine similarity between image A and image B with the n semantic part removed.

We define the two perturbed images as the pixel-wise multiplication of the images and the relative semantic section mask of the same size with values between 0 and 1.

$$A' = A \cdot M^{(A,n)} \quad \text{and} \quad B' = B \cdot M^{(B,n)} \quad (5)$$

The single removal operation is computed for all the semantic areas. For each mask, the value of the contribution map H0 is initialized with the C_n contribution associated with the mask:

$$H0_{(A,n)} = C_n \cdot M^{(A,n)} \quad (6)$$

The similarity map is defined as the sum of the negative and the positive contributions normalized by Equation 7 for all $n \in [1, m]$, to obtain $S0_A$.

$$H0_{(A,n)}^{\pm} = \begin{cases} \frac{H0_{(A,n)}}{\sum_{\substack{H0_{(A,m)} \geq 0 \\ H0_{(A,m)} < 0}} |H0_{(A,m)}|} & \text{if } H0_{(A,n)} \geq 0 \\ \frac{H0_{(A,n)}}{\sum_{\substack{H0_{(A,m)} < 0 \\ H0_{(A,m)} \geq 0}} |H0_{(A,m)}|} & \text{otherwise.} \end{cases} \quad (7)$$

$$S0_A = S0_A + (H0_{(A,n)}^+ + H0_{(A,n)}^-) \cdot M^{(A,n)}$$

We use the same Equations 6 and 7 to obtain $H0_{(B,n)}$, $H0_{(B,n)}^+$, $H0_{(B,n)}^-$ and $S0_B$. This means negative contributions are seen as dissimilar areas in the face image, while positive ones are similar. The algorithm 1 gives us the similarity maps $S0_A$ and $S0_B$ as a result of single removal.

3.1.4. Greedy Removal-S1

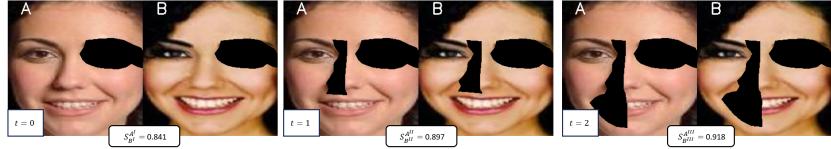


Figure 7: Greedy removal for image A and image B in n steps ($t = n - 1$), where $S_{B^{(n)}}^{A(n)}$ is the cosine similarity between the two images and n is the best part removed (Best_{M_A} and Best_{M_B}) at t -step.

The iterative approach of the greedy algorithm involves repeatedly performing a single removal procedure. In each iteration, the section of the face with the greatest impact is removed from images A and B . In particular, the initial images are represented as $A_0 = A$ and $B_0 = B$, and at each iteration, A_t and B_t are obtained by removing the principal parts of A_{t-1} and B_{t-1} , respectively. This means that at each iteration, the mask removed will be defined as the actual section mask sum with the previous best mask removed:

$$M^{(A_t)} = M^{(A_{t-1}, n)} + \text{Best}_{M_A} \quad \text{and} \quad M^{(B_t)} = M^{(B_{t-1}, n)} + \text{Best}_{M_B} \quad (8)$$

In greedy removal, calculating positive and negative contribution maps follows distinct procedures. We also use Equations 6 and 7 to obtain $H1_{(.,n)}$, $H1_{(.,n)}^+$, $H1_{(.,n)}^-$, $S1_A$ and $S1_B$. To be more concise, in algorithm 1, the presentation primarily focuses on calculating the negative contribution map $H1^-$. Indeed, in each iteration, the s_t value is set to 1. Consequently, the removed areas correspond to those exhibiting negative contribution, as the condition $s' < s_t$ dictates. Conversely, $H1^+$ is computed, setting s_t value to 0 at each iteration with the condition $s' > s_t$. In our example, the iteration stops when the maximum number of iterations is reached or when the score difference reaches a low enough point. In this case, that occurs at $t = 7$, where the score difference is only 0.009. After obtaining $H1^+$ and $H1^-$, the similarity map $S1$ is obtained following the equation 7.

3.1.5. Average Similarity map S_{AVG}

In subsection 3.1.3 and 3.1.4, the processes for determining similarity maps are outlined. Using single and greedy removal techniques makes it possible to assess the significance of each facial

Algorithm 1 Calculate H0 and H1

1: **Input:**

- 2: A -Face image A
- 3: B -Face image B
- 4: S_B^A -Initial Score
- 5: θ -Minimal increment allowed
- 6: t_{max} -Max number of iteration

7: $N, M \leftarrow \text{size}(A)$ ▷ height and width of face images
 8: $A_0 \leftarrow A$ ▷ initialization of the image
 9: $B_0 \leftarrow B$ ▷ initialization of the image
 10: $H0_A \leftarrow \text{zeros}(N, M)$ ▷ initialization of the maps
 11: $H0_B \leftarrow \text{zeros}(N, M)$ ▷ initialization of the maps
 12: $H1^-_A \leftarrow \text{zeros}(N, M)$ ▷ initialization of the maps
 13: $H1^-_B \leftarrow \text{zeros}(N, M)$ ▷ initialization of the maps
 14: $\text{Best}_{M_A} \leftarrow \text{zeros}(N, M)$ ▷ initialization of mask A
 15: $\text{Best}_{M_B} \leftarrow \text{zeros}(N, M)$ ▷ initialization of mask B
 16: $t = 0$ ▷ initialization of iteration counter
 17: $s_{t-1} \leftarrow S_B^A$ ▷ initial matching score
 18: $\Delta_s \leftarrow 1$ ▷ initialization of difference of scores
 19: **while** $t < t_{max}$ and $\Delta_s > \theta$ **do**
 20: $s_t \leftarrow 1$
 21: $t \leftarrow t + 1$
 22: **for** n in $\text{Face}_{\text{Sections}}$ **do**
 23: $M^{(A_t)} \leftarrow M^{(A_{t-1}, n)} + \text{Best}_{M_A}$
 24: $M^{(B_t)} \leftarrow M^{(B_{t-1}, n)} + \text{Best}_{M_B}$
 25: $A' = A_{t-1} \cdot M^{(A_t)}$
 26: $B' = B_{t-1} \cdot M^{(B_t)}$
 27: $s' \leftarrow S_{B'}^A$
 28: **if** $s' < s_t$ **then**
 29: $s_t = s'$
 30: $\text{Best}_{M_A} \leftarrow M^{(A_t)}$
 31: $\text{Best}_{M_B} \leftarrow M^{(B_t)}$
 32: $\text{Best}_{w_A} \leftarrow W(A, \text{Best}_{M_A})$
 33: $\text{Best}_{w_B} \leftarrow W(B, \text{Best}_{M_B})$
 34: $A_t \leftarrow A'$
 35: $B_t \leftarrow B'$
 36: $\Delta_{s_t} = s_{t-1} - s_t$
 37: **if** $t = 0$ **then**
 38: **for** n in $\text{Faces}_{\text{Sections}}$ **do**
 39: $C_n = \Delta_{s_t} \cdot \hat{W}_{(A, B)_n}$
 40: $H0_A[M^{(A, n)} = 1] \leftarrow C_n$
 41: $H0_B[M^{(B, n)} = 1] \leftarrow C_n$
 42: $C_{\text{Best}} \leftarrow \Delta_{s_t} \cdot \hat{W}_{(A, B)_{\text{Best}}}$
 43: $H1^-_A[\text{Best}_{M_A} = 1] \leftarrow C_{\text{Best}}$
 44: $H1^-_B[\text{Best}_{M_B} = 1] \leftarrow C_{\text{Best}}$

 45: **Output:** $H0_A, H1^-_A, H0_B, H1^-_B$

feature individually or in conjunction with others. Considering this, analyzing an average map can provide valuable comprehension of the significance of each facial feature. Incorporating this information within an average similarity map of maps S_0 (Single removal) and S_1 (Greedy removal) which is called S_{AVG} aligns with the notion that humans perceive and interpret faces in a relational/configurational manner [28] (see Figure 3).

4. Experimental results

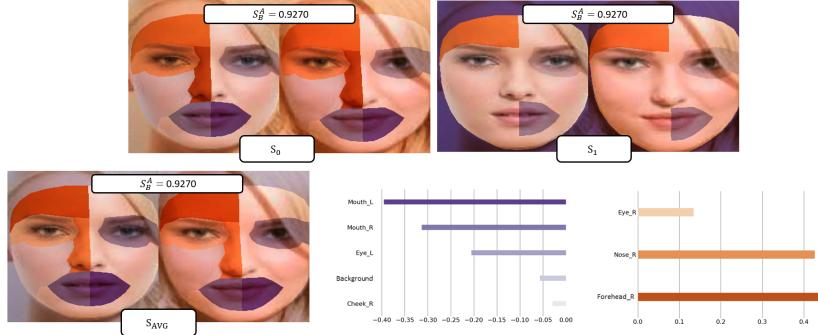


Figure 8: Similarity maps for each algorithm proposed in the case of VGGface2. Respectively S_0 is the output of the single removal, S_1 is the greedy removal ones, and S_{AVG} is the average map generated from S_0 , and S_1 . The plot chart considers the contribution values (C_n) for each section in the perturbation.

This section shows the experimental results for a selected number of samples extracted by the CelebA dataset [35] and tested for the FaceNet [37] model trained on Casia-WebFace [33] and VGGfaces2 [34]. In Figure 8, we show the output generated by the proposed method. It comprises three maps: the initial single removal map S_0 , the greedy removal map S_1 , and the ultimate average map S_{AVG} . The visualization uses orange to represent semantic areas that are similar, while purple indicates differences in facial features. After the concept extraction, a group of semantic areas is selected based on their importance. The table 1 displays the n-selected semantic areas ranked by their importance. In our study $n = 8$, this number can be changed as needed.

Table 1

Concept extraction output of each model's top n semantic areas (n=8). Area names are abbreviated with the initial or the first two letters (i.e., E = eye). R and L are associated with the right and left sides.

VGGFace2	"B", "CHE _R ", "MOL", "E _R ", "MO _R ", "N _R ", "F _R ", "E _L "
Casia Net	"B", "E _R ", "M _R ", "M _L ", "E _L ", "CHE _R ", "F _R ", "CHE _L "

Figure 8 also features a table showcasing the sections of the face categorized as similar (orange) and dissimilar (purple), along with their respective contribution values to the final similarity map. We will focus on analyzing the mean map S_{AVG} , which utilizes the same color scale. Regarding the nature of masking applied during perturbation, we investigated how it impacted the algorithm's output. In figure 9, we present two distinct case studies for both

models. The examined masking types encompass black masking, random noise masking, and white masking. Upon observing the images, it becomes evident that, in general, there is minimal sensitivity to the type of masking, especially between black and white masking. The most notable deviation is associated with random noise masking, although this divergence remains relatively modest. The maps reported in this study exclusively employ black masking. Figure

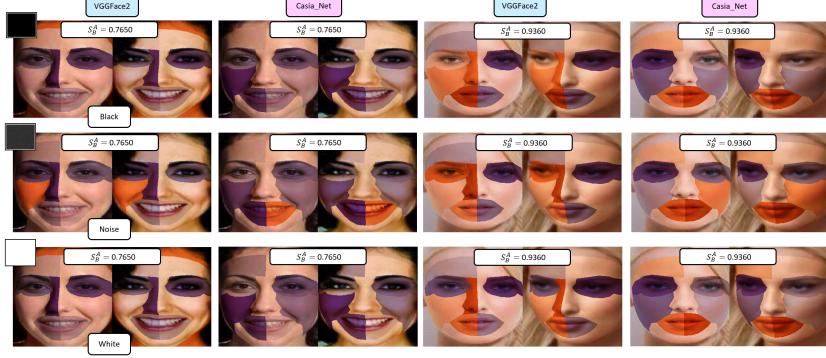


Figure 9: The S_{AVG} map for pair examples from the CelebA dataset, generated using different patch coloring for the models VGGface2 and CasiaNet. Orange hues denote similar facial regions, while purple highlights dissimilar ones.

10 presents several instances of the algorithm’s output for both tested models. Specifically, sections (a) and (c) demonstrate examples where facial comparisons are made between samples of the same individual, while sections (b) and (d) involve comparisons with imposters. Even when comparing faces of the same individual, certain areas are assessed as dissimilar, while conversely, when confronting imposters, not all areas are consistently regarded as dissimilar. The final score can offer additional insights by contextualizing which facial regions can be modified to influence the outcome.

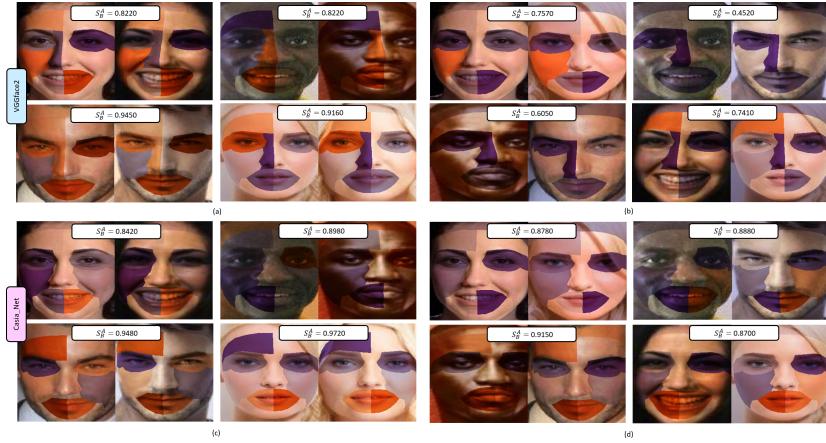


Figure 10: The S_{AVG} map for pair examples. In sections (a) and (c), the similarity maps are generated for genuine cases, while in sections (b) and (d) for impostor ones. The examples are generated from VGGface2 (a,d) and Casia Net (b,c).

4.1. Experiments with Cut-and-Paste Patches

We conducted a “Cut-and-Paste Patches” test to validate this outcome, as previously introduced by [16]. This experiment assesses whether replacing specific facial regions in one image with a corresponding region from another is effectively detected by our algorithm and described with high similarity in the similarity maps. We present the results in Figure 11. Specifically, in column (a), we display the average similarity map of the two original images. In column (b), one of the two images has been altered with a patch from the other (highlighted in green-yellow). Finally, in column (c), we present the resultant output. Overall, we observe that regions previously deemed dissimilar are now perceived as similar in the modified area, accompanied by an increase in the final score. Additionally, we notice instances where semantic areas change

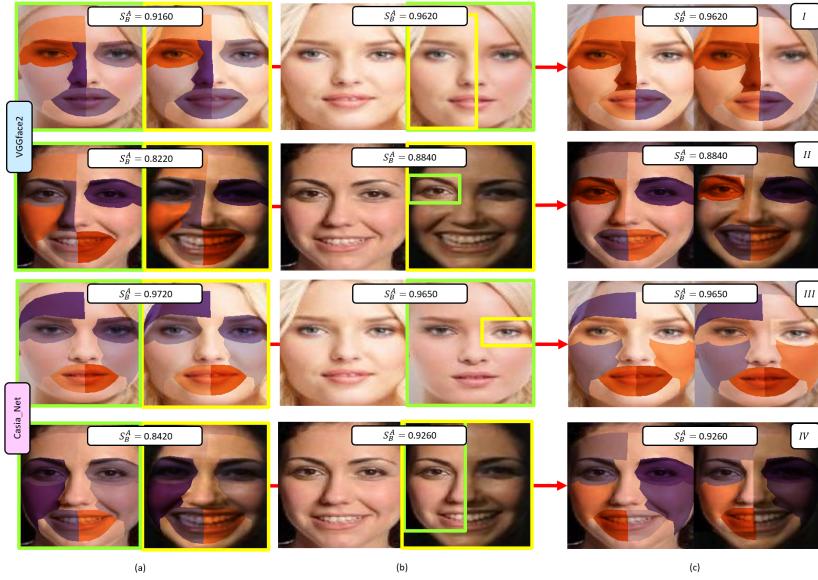


Figure 11: Cut-and-Paste patches test inspired by [38] for two samples output. In (a) the originals S_{AVG} , in (b) the pairs after applying the Cut-and-Paste Patches test, and in (c) the new average similarity maps.

in their contribution despite not being included in the modification patch. The explanation for this can be that the patches do not fit the exact dimensions as the semantic areas, and in some cases, a rectangular patch, mainly centered on one point, may intersect multiple subsequently affected semantic areas. This observation underscores the sensitivity of the proposed method, particularly the segmentation carried out by Mediapipe, to facial regions. It is also noteworthy that certain areas change in color even when they have not been directly modified – for instance, the right eye in Case I, the left cheek in Case II, the left eye in Case III, and in Case IV, the patch is not entirely recognized as similar. This discrepancy can be attributed to the fact that while the test follows a part-based approach, network models tend to perceive faces holistically, implying that altering a specific patch may lead to a change in perception of the entire face and not just the modified area. This explanation aligns with the study of Jacob et al.[38], which demonstrated that models trained on various datasets with the Thatcher effect [39] internalize a

holistic perception of faces. Moreover, it is essential to note that the maps under consideration focus solely on the most influential areas, albeit their influence on the final score is limited.

4.2. Method limitations

While Mediapipe offers valuable tools for semantically segmenting facial features, it displays a notable sensitivity to variations in facial orientation. Substantial deviations in facial pose result in increasingly dissimilar masks, leading to proportionally divergent contributions. When the masks exhibit high similarity, the simultaneous occlusion method gains coherence as it conceals identical portions of the image. Another limitation arises when comparing a profiled face with a frontal one. In such instances, Mediapipe can still identify facial features; however, the application of occlusion to both profiles loses its contextual relevance, rendering the affected areas visually less comprehensible. Consequently, the most suitable application of the method pertains to front-facing subjects with poses as closely aligned as possible.

5. Conclusion and Future directions

In this paper, we have initiated an effort to bridge the gap between computer and human vision, with the primary goal of improving the interpretability of facial verification algorithms. We sought to gain insight into how machines perceive the semantic aspects of human faces during verification, ultimately aligning the system’s output score more closely with human reasoning.

We employed the Mediapipe tool to identify distinct semantic regions on the human face to achieve this. These regions, representing human-conceptual knowledge, provided a comprehensive view of the critical concepts for our models. Leveraging this knowledge, we selected a subset of the most significant semantic areas for the models. We also introduced a perturbation algorithm that generated similarity maps, revealing how the models under examination perceived these concepts as either *similar* or *dissimilar*.

By contextualizing the system’s output score, we can align it more closely with human reasoning. However, it is essential to note that our work is currently limited to experimentation. As a result, future research directions could include exploring different segmentation methods, conducting experiments across diverse models, comparing various methods to ours, or adapting them to our approach. Additionally, including a user evaluation component could further validate and enhance the effectiveness of our work.

References

- [1] G. Alfarsi, J. Jabbar, R. M. Tawafak, A. Alsidiri, M. Alsinani, Techniques for face verification: Literature review, in: 2019 International Arab Conference on Information Technology (ACIT), IEEE, 2019, pp. 107–112.
- [2] J. Lynch, Face off: Law enforcement use of face recognition technology, Available at SSRN 3909038 (2020).
- [3] J. S. del Rio, D. Moctezuma, C. Conde, I. M. de Diego, E. Cabello, Automated border control e-gates and facial recognition systems, computers & security 62 (2016) 49–72.

- [4] D. J. Robertson, R. S. Kramer, A. M. Burton, Face averages enhance user recognition for smartphone security, *PloS one* 10 (2015) e0119460.
- [5] D. Zhang, N. Maslej, E. Brynjolfsson, J. Etchemendy, T. Lyons, J. Manyika, H. Ngo, J. C. Niebles, M. Sellitto, E. Sakhaei, et al., The ai index 2022 annual report. ai index steering committee, Stanford Institute for Human-Centered AI, Stanford University (2022) 123.
- [6] A. Olteanu, J. Garcia-Gathright, M. de Rijke, M. D. Ekstrand, A. Roegiest, A. Lipani, A. Beutel, A. Olteanu, A. Lucic, A.-A. Stoica, et al., Facts-ir: fairness, accountability, confidentiality, transparency, and safety in information retrieval, in: ACM SIGIR Forum, volume 53, ACM New York, NY, USA, 2021, pp. 20–43.
- [7] C. Davide, Can we open the black box of ai, *Nature News* 538 (2016) 20.
- [8] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, *ACM computing surveys (CSUR)* 51 (2018) 1–42.
- [9] S. S. Kim, E. A. Watkins, O. Russakovsky, R. Fong, A. Monroy-Hernández, " help me help the ai": Understanding how explainability can support human-ai interaction, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–17.
- [10] P. Bommer, M. Kretschmer, A. Hedström, D. Bareeva, M. M. Höhne, Finding the right XAI method - A guide for the evaluation and ranking of explainable AI methods in climate science, *CoRR* abs/2303.00652 (2023).
- [11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.
- [12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.
- [13] V. Petsiuk, A. Das, K. Saenko, Rise: Randomized input sampling for explanation of black-box models, *arXiv preprint arXiv:1806.07421* (2018).
- [14] B. Yin, L. Tran, H. Li, X. Shen, X. Liu, Towards interpretable face recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9348–9357.
- [15] D. Mery, B. Morris, On black-box explanation for face verification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 3418–3427.
- [16] M. Knoche, T. Teepe, S. Hörmann, G. Rigoll, Explainable model-agnostic similarity and confidence in face verification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 711–718.
- [17] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
- [18] M. Ribera, À. Lapedriza, Can we do better explanations? a proposal of user-centered explainable ai, in: IUI Workshops, 2019. URL: <https://api.semanticscholar.org/CorpusID:84832474>.
- [19] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, R. Sayres, Interpretability beyond feature attribution: Quantitative testing with Concept Activation Vectors (TCAV), in: 35th International Conference on Machine Learning (ICML), 2018, pp. 2668–2677.
- [20] Q. Zhang, R. Cao, F. Shi, Y. N. Wu, S.-C. Zhu, Interpreting cnn knowledge via an explanatory

- graph, in: 32nd AAAI Conference on Artificial Intelligence (AAAI), 2018, pp. 4454–4463.
- [21] A. Ghorbani, J. Wexler, J. Z. Y. B. Kim, Towards automatic concept-based explanations, in: Advances in Neural Information Processing Systems, volume 32, 2019.
 - [22] R. Tan, L. Gao, N. Khan, L. Guan, Interpretable artificial intelligence through locality guided neural networks, *Neural Networks* 155 (2022) 58–73.
 - [23] W. Zhang, B. Y. Lim, Towards relatable explainable ai with the perceptual process, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022.
 - [24] Perceptual processing / edited by Edward C. Carterette and Morton P. Friedman., *Handbook of perception* ; v. 9, Academic Press, New York, 1978.
 - [25] M. L. Matthews, Discrimination of identikit constructions of faces: Evidence for a dual processing strategy, *Perception & Psychophysics* 23 (1978) 153–161.
 - [26] G. Davies, H. Ellis, J. Shepherd, Cue saliency in faces as assessed by the ‘photofit’technique, *Perception* 6 (1977) 263–269.
 - [27] A. Iskra, H. Gabrijelčić Tomc, Eye-tracking analysis of face observing and face recognition, *Journal of Graphic Engineering and Design* 7 (2016) 5–11.
 - [28] G. Rhodes, Configural coding, expertise, and the right hemisphere advantage for face recognition, *Brain and cognition* 22 (1993) 19–41.
 - [29] P. Karczmarek, W. Pedrycz, A. Kiersztyn, P. Rutka, A study in facial features saliency in face recognition: An analytic hierarchy process approach, *Soft Comput.* 21 (2017) 7503–7517. doi:[10.1007/s00500-016-2305-9](https://doi.org/10.1007/s00500-016-2305-9).
 - [30] X. Li, Z. Yang, H. Wu, Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks, *IEEE Access* 8 (2020) 174922–174930.
 - [31] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), 2017, p. 4768–4777.
 - [32] J. Castro, D. Gómez, J. Tejada, Polynomial calculation of the shapley value based on sampling, *Computers & Operations Research* 36 (2009) 1726–1730.
 - [33] D. Yi, Z. Lei, S. Liao, S. Z. Li, Learning face representation from scratch, *arXiv* (2014).
 - [34] F. V. Massoli, G. Amato, F. Falchi, Cross-resolution learning for face recognition, *Image and Vision Computing* 99 (2020) 103927.
 - [35] H. Zhang, W. Chen, J. Tian, Y. Wang, Y. Jin, Show, attend and translate: Unpaired multi-domain image-to-image translation with visual attention (2018) 1–11.
 - [36] J. de Borda, Mémoire sur les élections au scrutin, *Histoire de L'Académie Royale des Sciences* 102 (1781) 657–665.
 - [37] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.
 - [38] G. Jacob, P. Rt, H. Katti, S. Arun, Qualitative similarities and differences in visual object representations between brains and deep networks, *Nature Communications* 12 (2021). doi:[10.1038/s41467-021-22078-3](https://doi.org/10.1038/s41467-021-22078-3).
 - [39] P. Thompson, Margaret thatcher: A new illusion, *Perception* 9 (1980) 483–484.