Comparing activation typicality and sparsity in a deep CNN to predict facial beauty

Sonia Tieo, Melvin Bardin, Nicolas M. Dibot, Roland Bertin-Johannet, Tamra C. Mendelson, William Puech, Julien Renoult

Comparing activation typicality and sparsity in a deep CNN to predict facial beauty

Sonia Tieo
CEFE, Univ Montpellier, CNRS, EPHE, IRD
Melvin Bardin
Terakalis
Roland Bertin-Johannet
CERCO, UMR5549
Nicolas Dibot
CEFE, Univ Montpellier, CNRS, EPHE, IRD
Tamra C. Mendelson
University of Maryland, Baltimore County
William Puech
LIRMM, Univ. Montpellier, CNRS
Julien P. Renoult
julien.renoult@cefe.cnrs.fr

CEFE, Univ Montpellier, CNRS, EPHE, IRD

Research Article

Keywords: beauty, fluency, statistical typicality, sparsity, CNNs, efficient coding

Posted Date: June 4th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4435236/v1

Additional Declarations: No competing interests reported.

Abstract

Processing fluency, which describes the subjective sensation of ease with which information is processed by the sensory systems and the brain, has become one of the most popular explanations of aesthetic appreciation and beauty. Two metrics have recently been proposed to model fluency: the sparsity of neuronal activation, characterizing the extent to which neurons in the brain are unequally activated by a stimulus, and the statistical typicality of activations, describing how well the encoding of a stimulus matches a reference representation of stimuli of the category to which it belongs. Using Convolutional Neural Networks (CNNs) as a model for the human visual system, this study compares the ability of these metrics to explain variation in facial attractiveness. Our findings show that the sparsity of neuronal activations is a more robust predictor of facial beauty than statistical typicality. Refining the reference representation to a single ethnicity or gender does not increase the explanatory power of statistical typicality. However, statistical typicality and sparsity predict facial beauty based on different layers of the CNNs, suggesting that they describe different neural mechanisms underlying fluency.

INTRODUCTION

Beauty holds significant influence across multiple aspects of human life. It shapes our perceptions, judgments (Batres & Shiramizu, 2020), preferences (Rhodes, 2006), and thereby guides our decision-making across diverse arenas, such as personal relationships (Rhodes et al., 2005) and consumer spending (Lee & Labroo, 2004). In the realm of cognitive science, processing fluency—defined as the sensation of ease of interpreting sensory information—has gained significant attention as a plausible determinant of people's evaluation of beauty (Reber et al., 2004). While fluency is currently subjectively measured through psychological experiments, few studies have attempted to model it, and none have compared the capabilities of existing models to predict beauty.

As a concept in aesthetic psychology, processing fluency provides a powerful explanation for a wide range of aesthetic inclinations, for both simple and complex stimuli. A preference for basic features such as symmetrical visual patterns and high contrasts is explicable by their effortless processing (Jacobsen et al., 2006; Reber et al., 2004, 1998). More complex stimuli, like fractal patterns, would be liked for their smooth processing as well, since the self-similarity of fractals at different scales makes these patterns highly predictive (Forsythe et al., 2011; Street et al., 2016). Fluency would also explain the attractiveness of prototypical representations (Winkielman et al., 2006). Prototypes are typified by familiar and easily discernible features, and thus prototype-like stimuli are processed with ease, enabling rapid and accurate categorization while enhancing memory retention. Regardless of their specific attributes, prototypes are consistently preferred across an array of stimuli, encompassing biological, inanimate, and abstract forms (Halberstadt & Rhodes, 2003; Winkielman et al., 2003; Winkielman et al., 2006).

Fluency, therefore, has a high explanatory power for beauty, and using this concept to predict beauty would have numerous technological applications, but also fundamental implications, for instance, in allowing us to study beauty in non-human animals (Renoult & Mendelson, 2019). However, the explanatory and predictive capacity of fluency is limited by the scarcity and current limits of studies aiming to model this concept. Early research in modeling processing fluency primarily focused on objective measures of feature repetitions in stimuli, such as symmetry, contrast, and self-similarity, for visual stimuli (Redies, 2007). While these metrics could provide valuable insights, they predominantly focus on the stimulus itself, assuming that the studied features ease information processing. Yet fluency is fundamentally rooted in the interaction between the stimulus and the perception of the beholder, and thus more accurate metrics should target features as they are processed by the human visual system.

A first step in modeling fluency is therefore to model the processing of features using a model of perception. The development of such models is uneven across the different sensory modalities, and is by far the most advanced for visual perception, which is the focus of this study (Mayer & Landwehr, 2018; Renoult & Mendelson, 2019). As demonstrated in numerous studies (Kriegeskorte, 2015; Lindsay, 2021), Deep Convolutional Neural Networks (CNNs) are powerful models of visual information processing, from low-level feature extraction to high-level semantic interpretation. Just as the human visual system processes information, CNNs begin by extracting simple features such as edges and contours in their initial layers. As the information flows through the network, increasingly complex features are recognized, which parallels the human visual system's ability to discern complex objects or scenes by combining simpler constituent elements. By studying neuronal activations within CNNs, which represent the response of different neurons of each layer to specific image inputs, we can thus gain valuable insights into how biological vision processes visual information.

The second step in modeling fluency is to choose a metric that characterizes the ease of information processing. Inspired by information theory applied to biological systems (H. B. Barlow, 1961), some authors have proposed to model fluency using the sparsity of neuronal activation (Redies, 2007; Renoult et al., 2016). Sparsity measures the concentration of activity in a subset of neurons. A sparse stimulus thus activates only a few neurons simultaneously, leading to a low-cost, efficient processing of information (Bruno A. Olshausen & Field, 2004). Using the sparsity of neuronal activations to estimate fluency fits the prediction of Winkielman et al. (Winkielman et al., 2012), that “fluent patterns should be represented by more extreme values of activation”, and “more differentiated states of theneurons”, meaning that the majority of neurons are not activated, but those that are, are highly activated. Previous studies have provided empirical evidence supporting the link between sparsity and beauty. For instance, using a model of information processing in the primary visual cortex, (Renoult et al., 2016) have shown a positive correlation between the sparsity of neuronal activations and the perceived attractiveness of female faces. Furthermore, sparsity has been identified as a robust predictor of face attractiveness compared to other factors such as body mass index, sexual dimorphism, averageness, and asymmetry (Holzleitner et al., 2019). More recently, one study evaluated the ability of the sparsity of neuronal activations within a CNN to explain variation in the beauty of faces and artistic paintings (Dibot et al., 2023). The authors showed that sparsity alone could explain up to 28% of the variance in beauty scores.

Another metric of fluency proposed in the literature is statistical typicality. Typicality describes the extent to which a stimulus aligns with an average representation. It is based on the underlying assumption that individuals form mental representations of averageness for various categories based on their past experiences and exposure to stimuli. Typicality is thus closely related to familiarity, which has been a well-studied factor influencing fluency in psychological studies (Reber et al., 1998). Ryali et al. (Ryali et al., 2020) demonstrated that the attractiveness of a face is partly explained by its statistical

typicality, defined as the likelihood of the face image relative to an internal representation of the face distribution for a given category of faces. The authors used the Active Appearance Model (AAM) as a model of face perception, which is built from features describing the shape and texture of faces. They then showed that the attractiveness of a given face can be predicted from its likelihood estimated from the distribution of faces of the same gender.

Here, we first propose a new model of fluency that applies the statistical typicality metric to convolutional neural networks (CNNs). More precisely, for each layer of the CNN, the method estimates the likelihood of a stimulus encoding given a reference distribution of encoded features. The method thus extends previous applications of statistical typicality ((Ryali et al., 2020); see also (Brielmann & Dayan, 2022)): our model of perception describes visual information processing as it operates throughout the retina and the ventral stream of the human visual system, and it is not specific to one visual domain (as is, e.g., the AMM model). Second, we aim to compare the ability of sparsity and statistical typicality computed in CNN layers to explain variation in the attractiveness of human faces. Attractiveness is arguably the strongest determinant of facial beauty, to a point that both terms are generally used interchangeably in the psychological literature (Rhodes, 2006). For this purpose, we used the publicly available Chicago Face Dataset (Ma et al., 2015), a comprehensive collection of face images with empirical scores of attractiveness. We encoded each face with a CNN, calculated its sparsity and likelihood (typicality) at each layer, and trained two regression models (one with sparsity, the other with likelihood) to predict attractiveness. Third, we examine the extent to which the ability of statistical typicality to explain facial attractiveness is influenced by the choice of the reference distribution. Specifically, we investigate whether specializing the reference distribution to include only faces of a single gender and/or a single ethnic group improves the ability of likelihood to predict attractiveness.

MATERIAL AND METHODS

1. Materials

We model the fluency of processing portrait images from Chicago Face Dataset (CFD hereafter; (Ma et al., 2015)). This dataset contains standardized photographic portraits (i.e., frontal view and standardized attire) of individuals aged 17 to 65, spanning a variety of ethnic backgrounds, including East Asian, Black, Hispanic, and White, with balanced representation across genders and ages. We use a subset of CFD, keeping only the 597 portraits depicting a neutral facial expression. This dataset also includes ratings of attractiveness, evaluated by independent judges. Each portrait is associated with one mean score of attractiveness.

2. Modeling face processing using CNNs

To model the visual processing of faces, we compare two CNNs, VGG16 (Simonyan & Zisserman, 2014) and VGGFace (Parkhi et al., 2015), that have been pre-trained on the ImageNet and VGGFace datasets, respectively. VGG16 includes 13 convolutions and two fully connected layers. ImageNet is a large dataset of 14 million images depicting about 20,000 categories including people, plants, animals and human-made objects. VGG16 trained on such a large and varied dataset allows modeling a visual cortex that is not specialized to one specific task (Güçlü & van Gerven, 2015; Peterson et al., 2018). In contrast, VGGFace is a variant of VGG16 specifically tuned for face recognition. The VGGFace dataset includes images capturing variation in facial expressions, angles, and lighting conditions. The fine-tuning of VGGFace to this dataset allows the entire network to adapt to face-specific features across all layers.

3. Metrics of fluency

In a CNN, the output of a convolutional layer is a matrix of size *HxWxC* where each entry represents the activation of a "neuron". The dimension *C* describes the number of channels, each embodying a unique feature, such as a distinct contrast or edge orientation. *HxW* is termed a feature map, describing where a feature is present in the image; *H* and *W* are the height and the width of the feature map, respectively.

Sparsity

The first metric of fluency is activation sparsity, which measures the concentration of activity in specific neurons. We use the method described in detail in Dibot et al. (2023). Briefly, we quantify sparsity using the Gini index (Hurley & Rickard, 2008), by flattening the activation matrices into one-dimensional vectors and sorting the vectors in ascending order. The Gini index is then computed as:

$$Gini = \frac{\sum_{i=1}^{n} (2i - n - 1) \cdot x_i}{n \cdot \sum_{i=1}^{n} x_i}$$

where *n* is the total number of activations in the layer (vector length) and x_i is the activation value at index *i*. Higher Gini indices indicate a higher degree of sparsity, reflecting a more selective activation.

Statistical typicality

The second metric of fluency is statistical typicality, presented here for the first time. It is calculated in three steps. The first step reduces the dimensionality of the *HxWxC* matrices. As the images traverse through a pre-trained network, a multitude of feature maps are generated from the different layers. Some of these feature maps have an extensive number of activations (for instance, the first convolutional layer of VGG16 yielded 3,211,264 activations), rendering the estimation of statistical typicality computationally intractable. We compare three strategies of dimensionality reduction. In the first strategy, we perform one Principal Component Analysis (PCA) per layer ("layer-wise PCA"), thus considering all activations (ie, after flattening the *HxWxC* matrix). We keep the number of principal components allowing us to account for 80% of the variance in activations (Iigaya et al., 2023). This number varies across layers and datasets (between 28 and 622). In the second strategy, a PCA is applied individually to each feature map, preserving the components that allowed us to account for 80% of intra-map variance. The components from all feature maps within a specific layer are then concatenated and reduced further using a second PCA, again keeping the number of principal components accounting for 80% of the variance. In the third strategy, we first calculate the mean activation of each feature map and then concatenate all the means into a single vector for each layer. Despite potential benefits, in particular the high computing speed of the

third strategy, neither the second nor the third strategy improved our results compared to the first one. We thus present only results obtained with the first strategy, based on a layer-wise PCA.

The second step for measuring statistical typicality was to establish a reference distribution of encoded features. Following the approach proposed by Brielmann & Dayan (2022) and Ryali et al. (2020), reference distributions are represented by Probability Density Functions (PDFs; one PDF per layer) from the reduced (by the layer-wise PCA) activations of all encoded portraits of a reference dataset. We use FairFace as a reference dataset (Karkkainen & Joo, 2021), a collection of face portraits categorized according to ethnicity, gender, and age, with a balanced representation of the different categories, enabling us to build an unbiased model of statistical typicality (see Fig. 1 for the pipeline schema). PDFs are estimated from a balanced subset of randomly selected 1,000 portrait images from FairFace. This number is chosen to allow a comparison of results when changing the reference dataset (see *Influence of the reference dataset*).

We estimate each PDF with Gaussian Mixture Models (GMMs), using the principal components as the variables in the Gaussian models (Fig. 1). With their capacity to combine simple Gaussian density functions or "Gaussian components", GMMs are able to capture the complexity of the underlying distribution of facial features. We determine the optimal number of Gaussian components with the Bayesian Information Criterion (BIC), setting an upper limit to 20 components.

In the third step, we use PDFs calculated from the reference datasets (FairFace) to estimate the log-likelihood (LLH) of each image of the Chicago Face Database (CFD), for every layer of the CNN. This is done after projecting the activation values of each image of CFD onto the principal components calculated using the FairFace dataset (Fig. 1).

One limitation of GMMs is their sensitivity to singularities. Concretely, this means that one Gaussian component can be fitted onto a single outlier, thus inflating the LLH of images located close to this outlier in the feature space. To mitigate this potential issue, the LLH of an image is calculated as the median value obtained after 100 replications of the analysis (including the second and third steps described previously, but not the first, dimensionality reduction step, because the layer-wise PCA is deterministic and thus is performed only once; see Fig. 1). Using Bayesian Mixture Gaussian Models as an alternative to address the problem of singularities yield similar results but at much higher computational cost (results not shown).

4. Statistical analyses

Using the face portraits of CFD, we assess the ability of layer-wise sparsity and statistical typicality (LLHs) of activations calculated from all convolutional and fully connected layers (‘AllLayers’ models, see below) to predict facial attractiveness. We conduct a regression analysis with attractiveness as the response variable, and the LLHs or sparsity values of every layer as independent variables. Our first model involving sparsity can be expressed as:

$$Attractiveness \sim \sum_{l=1}^{N} \left(a_l \cdot \mathrm{Sparsity}_{\mathrm{layer}_l} \right)$$

with *l* varying from 1 to N (the number of layers for VGG or VGG16), $\mathrm{Sparsity}_{\mathrm{layer}_l}$ indicating the sparsity of layer *l* and a_l the regression coefficient.

We similarly define our model for statistical typicality as:

$$Attractiveness \sim \sum_{l=1}^{N} \left(b_l \cdot LLH_{\mathrm{layer}_l} \right)$$

with *l* varying from 1 to N, LLH_{layer_l} representing each layer's *LLH* and b_l the regression coefficient.

We employ ridge regression models, rather than classical linear regression models, to address the inherent collinearity among layers, due to the fact that layer outputs are inputs of the following layers. Ridge regression effectively handles multicollinearity and overfitting by applying a regularization that balances model complexity and generalization (Hoerl & Kennard, 2000). We perform ridge regressions using the *glmnet* package in R and a 10-fold cross-validation repeated 10 times. For ridge regression (α = 1), we explore a range of penalty values (λ), spanning from 10^2 to 10^{-4}. Both the predictors and the response are centered and scaled prior to the analysis. The coefficient of determination, R^2, is the explained variance of attractiveness.

In addition to our primary focus on 'AllLayers', we delved into the distinct subsets of the VGGFace neural network layers. These subsets include 'LastConvLayer', 'PoolingLayers', 'ShallowLayers', 'MiddleLayers', and 'DeepLayers'. Each subset represents a specific arrangement of layers, with breakdowns detailed in supplementary information Figure S1. By analyzing these subsets, we aim to determine how different layers of the neural network contribute to predict facial attractiveness.

5. Influence of the reference dataset

Previous research suggests that the categorization of faces according to gender and ethnicity can impact perceived attractiveness (Potter & Corneille, 2008; Ryali & Yu, 2018). We therefore investigate if a more specialized reference Probability Density Function (PDF) — built using a reference dataset tailored more specifically to a particular gender, ethnic group, or a combination of the two — could enhance the ability of statistical typicality to explain variation in attractiveness.

We use the gender and ethnic categories of FairFace to build 15 reference datasets, each containing exactly 1,000 images (see details in Table 1). These 15 datasets can be categorized into four types of reference datasets differing in the level of specialization: “all”, “ethnicity”, “gender” and “ethnicity x gender”. We consider that “ethnicity x gender” represents a more specialized reference dataset than “ethnicity” and “gender”, which are themselves more specialized than “all”. Ethnic and gender categories are kept balanced within the “all”, “ethnicity” and “gender” reference datasets. As previously, we estimate PDFs (one per layer) for each reference dataset. Then, for each test image (representing one combination of gender and ethnic category from the CFD), we calculate its

LLHs with respect to increasingly narrow reference datasets from FairFace. For Asian male faces of the CFD dataset, for example, we calculate LLHs considering PDFs estimated from a reference dataset of 1,000 portraits of FairFace depicting either i) individuals of both genders and all (Asian, Black, White, Latino) ethnic groups (“all” reference dataset), ii) all Asian males and females (“ethnicity” reference dataset), iii) males of all ethnic groups (“gender” reference dataset), and iv) Asian males only (“ethnicity x gender” reference dataset) (Fig. 2).

To mitigate the problem of having a different number of images between gender and ethnic categories in CFD (R^2 is influenced by the number of observations, and adjusted-R^2 cannot be calculated in a ridge regression), we randomly sample 52 images from each category, equivalent to the minimum number of images present in any category. This sampling is repeated 20 times.

For each reference dataset, we thus obtain a set of LLHs (one per layer), which are as previously used in a ridge regression model to explain facial attractiveness. However, to address the challenge of having a large number of regressors relative to the number of observations (only 52 observations), rather than using raw LLHs values in the regression models, we reduce their number using a PCA. The first 5 principal components (PCs) cumulatively explained between 92–98% of variance. We calculate R^2 score for each ridge regression, using cross-validation as above. To test the hypothesis that a more specialized reference dataset leads to higher R^2 values, we then used a Generalized Linear Mixed Model (GLMM) with the R^2 score (derived from the mean of the 20 repetitions) as the response variable and the type of reference specialization (categorical, four levels, see Table 1) as the explanatory variable. The categories of “ethnicity x gender” within the CFD dataset (CFD_Categories_“ethnicity x gender”) are included as a random effect (categorical variable, eight levels, see Table 1) in the model. The model is thus expressed as:

$$R\square \sim PC1\,(LLH) + PC2\,(LLH) + PC3\,(LLH) + PC4\,(LLH) + PC5\,(LLH) + ReferenceSpecializationType + 1|CFD\_Categories\_“$$

We fit the model using lme4 in R.

Table 1

Description of Reference datasets built from FairFace dataset. Each reference dataset includes 1,000 images randomly sampled from the entire FairFace dataset (FaireFace_All), or from a subset of a single gender, a single ethnicity or a single gender of a single ethnicity. The column “Genders & ethnic groups" indicates the composition of the subset.

Genders & ethnic groups	Reference specialization type	Reference dataset identifier
all	all	FairFace_All
Asian, both genders	ethnicity	FairFace_A
Black, both genders	ethnicity	FairFace_B
Latino, both genders	ethnicity	FairFace_L
White, both genders	ethnicity	FairFace_W
Females, all ethnicity	gender	FairFace_F
Males, all ethnicity	gender	FairFace_M
Asian females	ethnicity x gender	FairFace_AF
Black females	ethnicity x gender	FairFace_BF
Latino females	ethnicity x gender	FairFace_LF
White females	ethnicity x gender	FairFace_WF
Asian males	ethnicity x gender	FairFace_AM
Black males	ethnicity x gender	FairFace_BM
Latino males	ethnicity x gender	FairFace_LM
White males	ethnicity x gender	FairFace_WM

Table 2

Description of the subsets of the Chicago Face Dataset test set. CFD is split into subsets including a single gender, a single ethnicity or a single gender of a single ethnicity (column “Genders & ethnic groups”). Each subset of CFD is analyzed with each reference dataset indicated in the column “Reference datasets compared”.

Genders & ethnic groups	Test dataset identifier	Number of images	Reference datasets compared
All	CFD_All	597	FairFace_ALL
Asian females	CFD_AF	57	FairFace_All/FairFace_A/FairFace_F/FairFace_AF
Black females	CFD_BF	104	FairFace_All/FairFace_B/FairFace_F/FairFace_BF
Latino females	CFD_LF	56	FairFace_All/FairFace_L/FairFace_F/FairFace_LF
White females	CFD_WF	90	FairFace_All/FairFace_W/FairFace_F/FairFace_WF
Asian males	CFD_AM	52	FairFace_All/FairFace_A/FairFace_M/FairFace_AM
Black males	CFD_BM	93	FairFace_All/FairFace_B/FairFace_M/FairFace_BM
Latino males	CFD_LM	52	FairFace_All/FairFace_L/FairFace_M/FairFace_LM
White males	CFD_WM	93	FairFace_All/FairFace_W/FairFace_M/FairFace_WM

RESULTS

Comparing statistical typicality to sparsity

We first evaluate the ability of sparsity of neuronal activations at each layer of the CNN to explain variation in facial attractiveness. Sparsity accounts for 27% and 23% of the variance in attractiveness with VGG16 and VGGFace, respectively (Fig. 3).

We then evaluate the ability of statistical typicality, measured at each layer of the CNN, to explain variation in facial attractiveness. Using portrait images of FairFace (FairFace_All in Table 1) as a reference dataset to calculate the LLHs of images of the CFD dataset (CFD_All in Table 2), we find that statistical typicality explains 8% of variance in facial attractiveness (R^2) with VGG16, and 11% with VGGFace (Fig. 3). These results indicate that the variation in the statistical typicality of neuronal activations triggered by facial features accounts for less variation in attractiveness compared to the sparsity of these activations.

We then investigate whether we could increase the explanatory capacities of statistical typicality by including LLHs from a subset of VGGFace layers, the CNN that yielded the highest R^2 score for this fluency metric, rather than considering all layers. We consider four subsets of layers: regression models with 'LastConvLayer' include the last convolutional layer of each block, 'PoolingLayers' models includes all pooling layers, 'ShallowLayers', 'MiddleLayers' and 'DeepLayers' models include early, mid-tier and the deeper layers of the network, respectively. The specific layers included in each of these subsets are detailed in Figure S1. None of these subsets yield at R^2 score surpassing that of sparsity. More precisely, we find that the fraction of explained variance is lower compared to when considering all layers, except with the 'DeepLayers' subset that yield similar performances (11% of explained variance of attractiveness; Figure S1).

Influence of reference dataset specialization

In the previous analysis of statistical typicality, we find a slightly higher explanatory power when using VGGFace (with weights tuned using VGGFace dataset) compared to VGG16 (with weights tuned using ImageNet). This could be due to the specialization of VGGFace to process faces, leading to PDFs that are more tightly tuned to facial features and thus to more meaningful values of LLH. We thus assess whether the statistical typicality metric would be more predictive when specializing the PDFs even further, such that LLHs are calculated in reference to one gender or one ethnic group only, or even one gender of one ethnic group, rather than all faces considered together.

To investigate the role of specializing the reference datasets and PDFs further, we perform a regression model with the R^2 score of the 15 models with different levels of specialization (Figure S2) as a response variable. When analyzing the different levels of the categorical variable “Reference specialization type”, we find that 'ethnicity x gender' (estimate = -0.03, 95% CI [-0.06, 0], p = 0.027; Table 2) significantly but negatively influences the explanatory power of statistical typicality. This result indicates that specializing the reference dataset and associated PDFs does not increase the ability of statistical typicality to explain variation in facial attractiveness. On the contrary, we obtain the best performance when using the 'All' reference dataset, that is, when images are sampled across all genders and ethnic groups of the FairFace dataset. Importantly, variation in R^2 is not due to differences in sample size, which are kept constant across datasets (52 images).

Table 2

Influence of reference dataset specialization on facial attractiveness prediction. Results from GLMM analysis exploring how different reference specialization types – 'ethnicity', 'gender', and 'ethnicity x gender' – influence the accuracy of predicting facial attractiveness. Fixed effects include the number of photos from CFD categories, while random effects encompass the eight categories "ethnicity x gender" from CFD images (as detailed in Table 1). The 'All' reference specialization type, with diverse images across genders and ethnicities, yields significantly better predictions of facial attractiveness than 'ethnicity' and 'ethnicity x gender' reference specialization types.

Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.32	0.28–0.35	**< 0.001**
Reference specialization type [ethnicity] (ref: All)	-0.02	-0.05–0.01	0.135
Reference specialization type [gender] (ref: All)	-0.02	-0.05–0.01	0.112
Reference specialization type [ethnicity x gender] (ref:All)	-0.03	-0.06–0.00	**0.027**
Random effects			
N CFD_Categories_"ethnicity x gender"	8		
Observations	32		
Marginal R^2/ Conditional R^2	0.044 / 0.593		

DISCUSSION

The objectives of this study were to first present a new fluency-based predictive model of facial attractiveness relying on statistical typicality of neuronal activations, and then to compare it to another metric relying on the sparsity of activations. While statistical typicality can contribute to predict facial beauty, we found that the sparsity of neuronal activations is a better predictor.

Statistical typicality predicts facial beauty

Our CNN-derived measure of statistical typicality was able predict some of the variation in attractiveness scores associated with the faces in the CFD dataset (R^2 = 0.11). Interestingly, the predictive power was similar when considering all layers or only the deepest (i.e. last) layers of the CNNs, and adding information from shallower layers did not increase the R^2 score further. In a CNN, the deep layers encode complex patterns and their arrangement at the largest spatial level; for faces, this means the shape and coloration of entire facial elements (i.e. the nose, the mouth, the eyes) and their relative position in the face (Khan et al., 2020). In other words, statistical typicality would be linked to beauty during configural perception of a face. This result is consistent with that of a previous study by Ryali et al. (2020), which inspired our work, that analyzed the relationship between attractiveness and statistical typicality of CFD faces within the Active Appearance Model (AAM) of face representation. The AAM model integrates both low (texture) and high-level (global form) information, the latter of which is analogous to the deeper layers of our CNNs. The authors found a strong correlation (r = 0.386, corresponding to R^2 = 0.15) between attractiveness ratings and the LLH of the stimuli, which is close to the R^2 = 0.11 obtained with our CNNs. When considering statistical typicality only, thus, a CNN does not seem to provide advantage over the AMM to encode faces.

Our statistical typicality-based fluency model also exhibits similarities with the model of aesthetics proposed by Brielmann et al. (2022), and experimentally tested in Brielmann et al. (2024).In their experiment, the authors also used VGG16 to model the encoding of visual stimuli, conducted a PCA to reduce the dimensionality of the encoding spaces, and estimated the Probability Density Function (PDF) and the log-likelihood (LLH) to describe statistical typicality. Their model assumes that aesthetic experience is determined by two types of rewards: the immediate sensory reward associated with fluency, and the value of learning. The immediate sensory reward is calculated through the LLH of a stimulus given an observer's system state (the PDF), similar to our model. The value of learning is computed as the change in the average likelihood of expected future stimuli. The authors demonstrated their model's ability to predict inter-individual variation in aesthetic judgments of visual stimuli. However, they also fund that, compared to the immediate sensory reward, the value of learning only moderately contributes to model performance, underscoring the primacy of fluency in aesthetic experience and the capacity of statistical typicality to model this fluency. The authors emphasize that the relative importance of the two rewards is likely to vary among individuals and stimuli, particularly based on prior knowledge of the stimuli, their motivation to learn, and the capacity of the stimuli to arouse curiosity. Nevertheless, we believe that these components and the reward of learning represent facets of aesthetic experience that are distinct from beauty. Indeed, several studies have shown that the evaluation of beauty and the pleasure of viewing reach their maximum at a glance (Brielmann et al., 2017; Locher et al., 2007). Beauty would thus describe the ease of information processing mostly during the bottom-up processing of information. As Francis Hutcheson (Hutcheson & Kivy, 1973) wrote about beauty: "pleasure does not arise from any knowledge of principles, proportions, causes or of the usefulness of the object; but strikes us at first with the idea of beauty".

No improvement when refining the reference distribution

Because statistical typicality relies on the category of stimuli being referenced, we replicated the analysis with different reference distributions, each describing a group of faces belonging to the same gender, ethnic group, or a combination of the two. Contrary to our expectations, refining the reference distributions (i.e., using PDFs that describe only one of these groups) did not improve the predictive performance of statistical typicality. This result may appear surprising, given those of previous studies showing that the brain interprets facial cues in the context of ethnic or gender-related subgroups (Kondo et al., 2013; Kramer et al., 2013; Levin, 1996; Ryali et al., 2020). For instance, (Potter & Corneille, 2008) found that computer-generated Caucasian and African-American faces are more attractive when they resemble the average features of their specific racial group. Yet, other studies also suggest that facial beauty could be evaluated in unique face space, a perceptual construct representing facial dissimilarity. A robust finding in psychology, for example, is “beauty-in-averageness”, which describes the fact that an average face, even of people with different ethnic origin, tend to be perceived as more attractive than the parental faces (Lewis, 2010; Rhodes et al., 2005).One explanation for the discrepancy between these two types of results is that statistical typicality can be computed in different sub-spaces within the face space depending on the perceptual tasks performed with these faces. For example, Ryali et al. (2020) and Halberstadt and Winkielman (2014) showed that preceding attractiveness evaluation with an ethnicity categorization task renders biracial faces less attractive. This result is well explained by the hypothesis that statistical typicality influences attractiveness: the racial categorization task embeds the attractiveness evaluation task along an axis where the face distribution appears obviously multimodal, making biracial faces atypical (because they lie between two distributions) and therefore less attractive. However, without this prior categorization task, Ryali et al. (2020) found that hybrid faces were more attractive, in according with the “beauty-in-averageness” hypothesis. These and our own results therefore suggest that, spontaneously, attractiveness evaluation occurs within a unique face space where mixed-race faces appear more typical of a generic perceptual category 'face' than ethnically typed faces.

A complementary explanation probably lies in the origin of raters. In the CFD database, a single attractiveness score is assigned to each face by pooling scores from multiple raters. All raters are of North American origin, and if the database does not provide this information, it can be reasonably assumed that the raters represent diverse ethnic backgrounds of each gender. Additionally, in the modern era, Westerners such as the raters in the CFD are regularly exposed to faces of various ethnicities and thus construct a relatively universal face space, with the prototype being a gender and ethnically composite face (Halberstadt & Rhodes, 2003; Lewis, 2010; Rhodes et al., 2005). To confirm these explanations, it would be interesting in the future to work with raters of different ethnic backgrounds living in societies that are not yet fully globalized. We would then predict that refining the reference distributions has an effect on the ability of statistical typicality to predict beauty ratings.

Sparsity outperforms statistical typicality

A second objective of this study was to compare statistical typicality with another proxy of fluency: the sparsity of neuronal (or feature) activations. Using the same dataset, and statistical analyses including the same number of explanatory variables for a fair comparison between metrics, we found that sparsity outperforms statistical typicality for predicting facial attractiveness.

To attempt to understand why sparsity surpasses typicality, one must analyze their links to underlying biological processes. Statistical typicality and sparsity are both related to fluency through the efficient coding theory. This theory posits that over the course of evolution, the brain has developed multiple strategies to process information in an economical way (H. Barlow, 2001; Chalk et al., 2018). One of these strategies is the sparse coding of information (B. A. Olshausen & Field, 1997; Bruno A. Olshausen & Field, 2004). Given that the majority of the brain's energy cost comes from the activation of neurons rather than their maintenance (Attwell & Laughlin, 2001), the brain minimizes the use of its metabolic resources by activating only a small proportion of a large number of neurons, all highly specialized, at any given time (Simoncelli & Olshausen, 2001). Sparsity is primarily enabled by tuning neurons to the signal features that are most frequently encountered, such as those that characterize natural environments, or that are important for the reproduction or survival of individuals (e.g., detection and recognition of partners or food; (Simoncelli & Olshausen, 2001)). As a result, stimuli that have a strong likelihood (statistically typical) also produce sparse encoding. Accordingly, we found that across the layers of our CNN, statistically typical faces are also sparsely encoded (mean Pearson correlation between sparsity and LLH: R = 0.4 for VGG, R = 0.15 for VGGFace).

However, the correlation is not maximal, explaining why the two metrics can account for facial beauty differently. As previously discussed, with statistical typicality we found that the fraction of explained variance was the highest for the deeper layers. This is in contrast with previous results on sparsity. Dibot et al. (2023) analyzed the contribution of the sparsity of individual CNN’s layers to facial attractiveness and found that the first layers of VGG16 explained most of the variance. Their result was consistent with those of Renoult et al. (2016), who showed that activation sparsity in a model of the primary visual cortex explains up to 17% of the variance. Indeed, in VGG16 the first convolutional layers have been shown to model feature extraction as it operates in the primary visual cortex (Lindsay, 2021). We thus suggest that different mechanisms of fluency contribute to facial attractiveness. In the early stages of information processing (modeled by the shallowest layers of a CNN), the sparsity of encoding local features such as the smoothness of skin texture would be the main driver of beauty, possibly because the small size of the neuron's receptor field and the ubiquity of the features they encode (e.g., abstract line contrasts) require the activation of many of these neurons, making sparsity a critical means of producing an economical neuronal code. In the later stages of information processing (modeled by the deeper layers of a CNN), efficiency in processing configurational information of the stimulus would have a greater influence on beauty, possibly because what determines fluency here would be the ability to recognize and memorize faces. Importantly, if they affect different stages of the information processing pathway, at some point sparsity and statistical typicality should be combined to generate fluency. Indeed, previous studies in psychology have shown that the ease of processing information at different stages of the visual system, as measured by detection time in the early stages and recognition performance in the later stages, triggers micro-experiences of fluency that aggregate into one global sensation of fluency (Reber et al., 2004; Wurtz et al., 2008). To understand better how sparsity and typicality differently and/or interactively influence fluency, further studies are now needed to correlate these metrics with empirical measures of fluency in detection and recognition tasks.

In conclusion, modeling fluency in perception involves capturing how a stimulus is processed within an observer's brain. In this study, we leveraged the potential of Convolutional Neural Networks (CNNs) to model information processing within the visual system and to probe the underlying neural mechanisms of fluency. We found that the sparsity of neuronal activation, which portrays the efficiency of neural information processing, appears to be a more powerful

determinant of beauty than statistical typicality. However, statistical typicality and sparsity predict facial beauty based on different layers of the CNNs, suggesting that they describe different neural mechanisms underlying fluency.

Declarations

STATEMENTS & DECLARATIONS

Funding

This study was funded by the Agence Nationale de la Recherche (ANR-20-CE02-0005-01), the National Science Foundation (NSF IOS 2026334) and by the Mission for Interdisciplinarity of the French National Center for Scientific Research (Programme Interne Blanc CNRS MITI 2023.1 – DEEPCOM project).

Competing Interests

The authors declare no competing interests.

Statements

The authors have no relevant financial or non-financial interests to disclose

Author Contributions

Sonia Tieo: Writing – Original draft, Methodology, Supervision, Formal Analysis, Software.

Tamra Mendelson: Conceptualization, Supervision, Funding Acquisition, Writing – Review & Editing.

Julien P. Renoult: Conceptualization, Supervision, Funding Acquisition, Writing – Review & Editing.

William Puech: Conceptualization, Supervision, Funding Acquisition, Writing – Review & Editing.

Melvin Bardin: Methodology, Formal Analysis, Software

Roland Bertin-Johannet: Methodology, Formal Analysis, Software

Nicolas Dibot: Methodology, Supervision

Ethics approval

This research only include simulation studies and therefore ethical approval was not required.

Consent to Participate

This doesn't apply because this research does not involve any participants.

Consent to Publish

This doesn't apply because this research does not involve any participants.

References

1. Attwell, D., & Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism: Official Journal of the International Society of Cerebral Blood Flow and Metabolism*, *21*(10), 1133–1145.
2. Barlow, H. (2001). Redundancy reduction revisited. *Network (Bristol England)*, *12*(3), 241–253.
3. Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. *Sensory Communication* (pp. 216–234). The MIT.
4. Batres, C., & Shiramizu, V. K. M. (2020). PSA001 Secondary Analysis: Examining the attractiveness halo effect across cultures. In *PsyArXiv*. https://doi.org/10.31234/osf.io/c7hf3.
5. Brielmann, A. A., Berentelg, M., & Dayan, P. (2024). Modelling individual aesthetic judgements over time. *Philosophical Transactions of the Royal Society of London Series B Biological Sciences*, *379*(1895), 20220414.
6. Brielmann, A. A., & Dayan, P. (2022). A computational model of aesthetic value. *Psychological Review*, *129*(6), 1319–1337.
7. Brielmann, A. A., Vale, L., & Pelli, D. G. (2017). Beauty at a glance: The feeling of beauty and the amplitude of pleasure are independent of stimulus duration. *Journal of Vision*, *17*(14), 9.
8. Chalk, M., Marre, O., & Tkačik, G. (2018). Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(1), 186–191.
9. Dibot, N., Tieo, S., Mendelson, T., Puech, W., & Renoult, J. (n.d.). Sparsity in an artificial neural network predicts beauty: towards a model of processing-based aesthetics. In *In prep*.

10. Forsythe, A., Nadal, M., Sheehy, N., Cela-Conde, C. J., & Sawey, M. (2011). Predicting beauty: fractal dimension and visual complexity in art. *British Journal of Psychology*, *102*(1), 49–70.
11. Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(27), 10005–10014.
12. Halberstadt, J., & Rhodes, G. (2003). It's not just average faces that are attractive: computer-manipulated averageness makes birds, fish, and automobiles attractive. *Psychonomic Bulletin & Review*, *10*(1), 149–156.
13. Hoerl, A. E., & Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics: A Journal of Statistics for the Physical Chemical and Engineering Sciences*, *42*(1), 80.
14. Holzleitner, I. J., Lee, A. J., Hahn, A. C., Kandrik, M., Bovet, J., Renoult, J. P., Simmons, D., Garrod, O., DeBruine, L. M., & Jones, B. C. (2019). Comparing theory-driven and data-driven attractiveness models using images of real women's faces. *Journal of Experimental Psychology Human Perception and Performance*, *45*(12), 1589–1595.
15. Hurley, N., & Rickard, S. (2008, October). Comparing measures of sparsity. *2008 IEEE Workshop on Machine Learning for Signal Processing*. 2008 IEEE Workshop on Machine Learning for Signal Processing (MLSP) (Formerly known as NNSP), Cancun, Mexico. https://doi.org/10.1109/mlsp.2008.4685455.
16. Hutcheson, F., & Kivy, P. (1973). *Francis Hutcheson: An inquiry concerning beauty, order, Harmony, design* (P. Kivy, Ed.) [PDF]. Kluwer Academic.
17. Iigaya, K., Yi, S., Wahle, I. A., Tanwisuth, S., Cross, L., & O'Doherty, J. P. (2023). Neural mechanisms underlying the hierarchical construction of perceived aesthetic value. *Nature Communications*, *14*(1), 127.
18. Jacobsen, T., Schubotz, R. I., Höfel, L., & Cramon, D. Y. (2006). Brain correlates of aesthetic judgment of beauty. *Neuroimage*, *29*(1). https://doi.org/10.1016/j.neuroimage.2005.07.010.
19. Karkkainen, K., & Joo, J. (2021, January). FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA. https://doi.org/10.1109/wacv48630.2021.00159.
20. Khan, K., Attique, M., Khan, R. U., Syed, I., & Chung, T. S. (2020). A multi-task framework for facial attributes classification through end-to-end face parsing and Deep Convolutional Neural Networks. *Sensors (Basel Switzerland)*, *20*(2), 328.
21. Kondo, A., Takahashi, K., & Watanabe, K. (2013). Influence of gender membership on sequential decisions of face attractiveness. *Attention Perception & Psychophysics*, *75*(7), 1347–1352.
22. Kramer, R. S. S., Jones, A. L., & Sharma, D. (2013). Sequential effects in judgements of attractiveness: the influences of face race and sex. *PloS One*, 8(12), e82226.
23. Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.
24. Lee, A. Y., & Labroo, A. A. (2004). The Effect of Conceptual and Perceptual Fluency on Brand Evaluation. *JMR Journal of Marketing Research*. https://doi.org/10.1509/jmkr.41.2.151.28665.
25. Levin, D. T. (1996). Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology Learning Memory and Cognition*, *22*(6), 1364–1382.
26. Lewis, M. B. (2010). Why are mixed-race people perceived as more attractive? *Perception*, *39*(1), 136–138.
27. Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031.
28. Locher, P., Krupinski, E. A., Mello-Thoms, C., & Nodine, C. F. (2007). Visual interest in pictorial art during an aesthetic experience. *Spatial Vision*, *21*(1–2), 55–77.
29. Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135.
30. Mayer, S., & Landwehr, J. R. (2018). Quantifying visual aesthetics based on processing fluency theory: Four algorithmic measures for antecedents of aesthetic preferences. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/ACA0000187.
31. Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research*, *37*(23), 3311–3325.
32. Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14*(4), 481–487.
33. Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). *Deep Face Recognition*. https://doi.org/10.5244/C.29.41.
34. Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and Improving) the Correspondence Between Deep Neural Networks and Human Representations. *Cognitive Science*, *42*(8), 2648–2669.
35. Potter, T., & Corneille, O. (2008). Locating attractiveness in the face space: faces are more attractive when closer to their group prototype. *Psychonomic Bulletin & Review*, *15*(3), 615–622.
36. Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology Inc*, *8*(4), 364–382.
37. Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science*, *9*(1), 45–48.
38. Redies, C. (2007). A universal model of esthetic perception based on the sensory coding of natural stimuli. *Spatial Vision*, *21*(1–2), 97–117.
39. Renoult, J. P., Bovet, J., & Raymond, M. (2016). Beauty is in the efficient coding of the beholder. *Royal Society Open Science*, *3*(3), 160027.

40. Renoult, J. P., & Mendelson, T. C. (2019). Processing bias: extending sensory drive to include efficacy and efficiency in information processing. *Proceedings. Biological Sciences*, *286*(1900), 20190165.
41. Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, *57*(1), 199–226.
42. Rhodes, G., Simmons, L. W., & Peters, M. (2005). Attractiveness and sexual behavior: Does attractiveness enhance mating success? *Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society*, *26*(2), 186–201.
43. Ryali, C. K., Goffin, S., Winkielman, P., & Yu, A. J. (2020). From likely to likable: The role of statistical typicality in human social assessment of faces. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(47), 29371–29380.
44. Ryali, C. K., & Yu, A. J. (2018). Beauty-in-averageness and its contextual modulations: A Bayesian statistical account. In *bioRxiv*. bioRxiv. https://doi.org/10.1101/360651.
45. Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216.
46. Simonyan, K., & Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. http://arxiv.org/abs/1409.1556.
47. Street, N., Forsythe, A. M., Reilly, R., Taylor, R., & Helmy, M. S. (2016). A Complex Story: Universal Preference vs. Individual Differences Shaping Aesthetic Response to Fractals Patterns. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00213.
48. Winkielman, P., Huber, D. E., Kavanagh, L., & Schwarz, N. (2012). *Fluency of consistency: When thoughts fit nicely and flow smoothly. Cognitive Consistency: A Fundamental Principle in Social Cognition*. 89–111.
49. Winkielman, P., Schwarz, N., Fazendeiro, T. A., & Reber, R. (2003). *The hedonic marking of processing fluency: Implications for evaluative judgment*. https://www.semanticscholar.org/paper/The-hedonic-marking-of-processing-fluency%3A-for-Winkielman-Schwarz/750bf4a9044a127106a89bad9f90c01741f6adad.
50. Winkielman, P., Halberstadt, J., Fazendeiro, T., & Catty, S. (2006). Prototypes are attractive because they are easy on the mind. *Psychological Science*, *17*(9), 799–806.
51. Wurtz, P., Reber, R., & Zimmermann, T. D. (2008). The feeling of fluent perception: a single experience from multiple asynchronous sources. *Consciousness and Cognition*, *17*(1), 171–184.

Figures

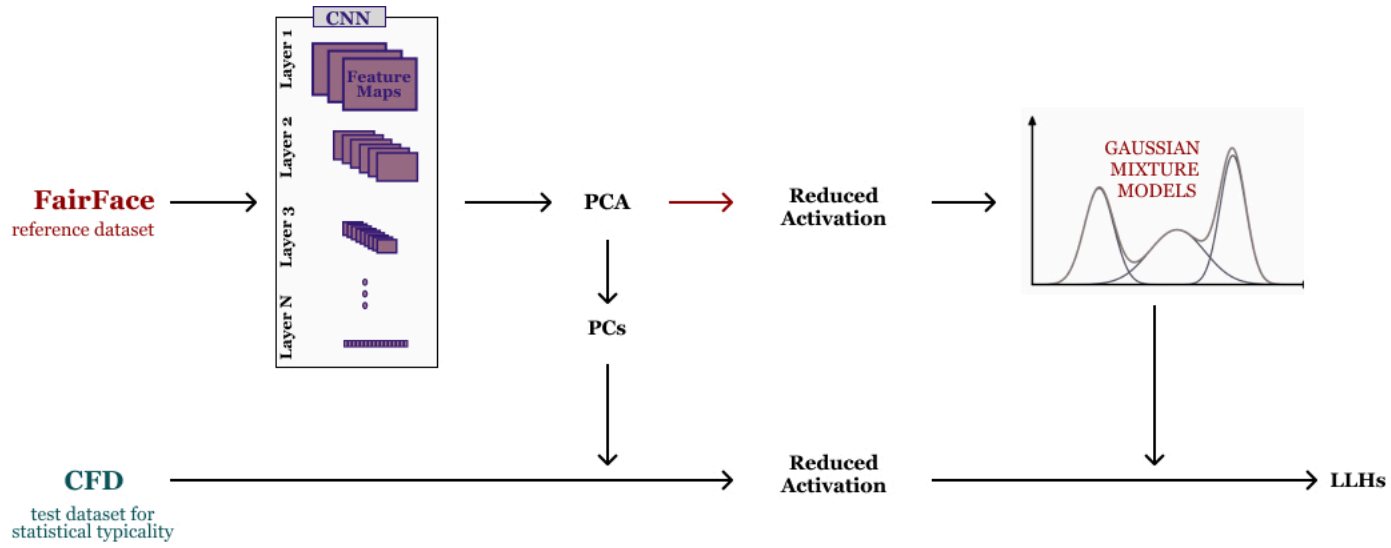

Figure 1

Pipeline for estimating the statistical typicality of faces. Images of the FairFace dataset (reference) are first encoded with a CNN (VGG16 or VGGFace). For each layer of the CNN, the dimensionality of the activation space describing the encodings is then reduced using a layer-wise principal component analysis (PCA), keeping only the principal components (PCs) explaining 80% of the variance. One probability density function (PDF) per layer is estimated using Gaussian Mixture Models (GMM) fitted onto the retained PCs. In a second step, images of the CFD dataset are encoded as in the first step, and their log-likelihood (LLH) calculated for each layer using the previously calculated PDFs. This entire process is repeated 100 times, and the statistical typicality is calculated as the median of all repetitions.

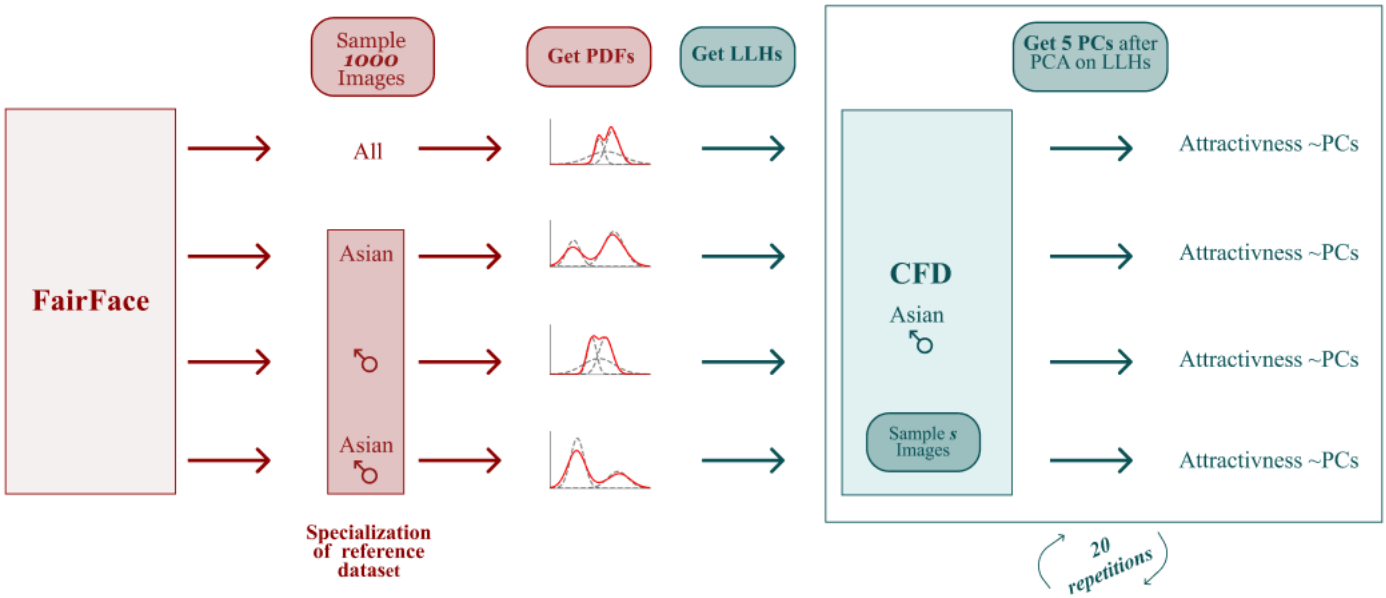

Figure 2

Log-Likelihood (LLH) calculation using different reference datasets - An example using the Asian males category. LLH values for Asian males in the Chicago Face Dataset are calculated using four types of reference dataset: 'all', 'Asian' (ethnicity-specific subset from FairFace), 'Males' (gender-specific subset from FairFace), and 'Asian Males' (ethnicity and gender-specific subset from FairFace). To address variation in the number of images for each gender and ethnic category within the CFD dataset (see Table 1), we randomly select a set of ***s=52***images from each category, which corresponds to the smallest category size. The derived LLH values undergo dimensionality reduction using PCA, resulting in the first 5 principal components (PCs). These 5 PCs were then incorporate into a ridge regression model trained to predict facial attractiveness. This sampling process was carried out 20 times to ensure consistency. The process delineated in the figure is representative of the method applied to all 15 categories (Table 1).

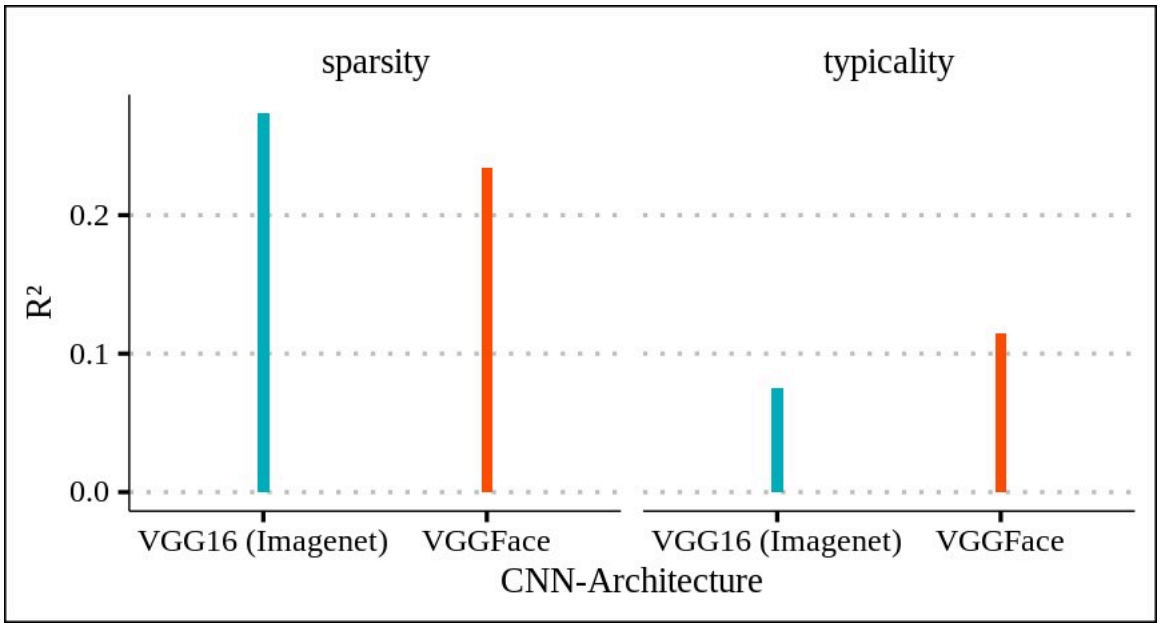

Figure 3

Comparison of explained variance of attractiveness (R^2) using sparsity and statistical typicality metrics derived from VGG16(Imagenet) and VGGFace architectures.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- suppmaterial.docx