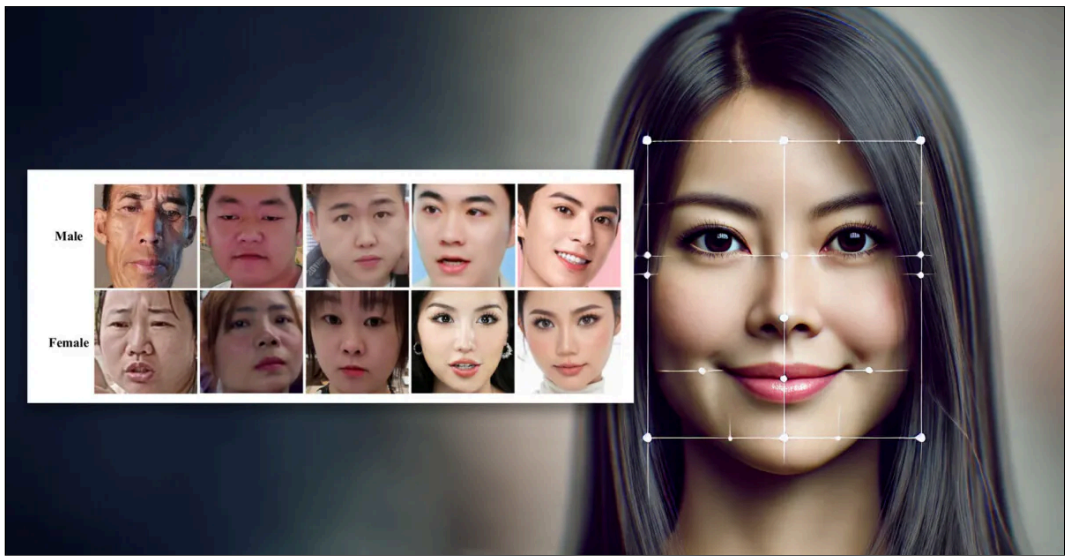
Estimating Facial Attractiveness Prediction for Livestreams

Originally published at https://www.unite.ai/estimating-facial-attractiveness-prediction-for-livestreams/



To date, Facial Attractiveness Prediction (FAP) has primarily been studied in the context of psychological research, in the beauty and cosmetics industry, and in the context of cosmetic surgery. It's a challenging field of study, since standards of beauty tend to be national rather than global.

This means that no single effective AI-based dataset is viable, because the mean averages obtained from sampling faces/ratings from all cultures would be very biased (where more populous nations would gain additional traction), else applicable to *no culture at all* (where the mean average of multiple races/ratings would equate to no actual race).

Instead, the challenge is to develop *conceptual methodologies* and workflows into which country or culture-specific data could be processed, to enable the development of effective per-region FAP models.

The use cases for FAP in beauty and psychological research are quite marginal, else industry-specific; therefore most of the datasets curated to date contain only limited data, or have not been published at all.

The easy availability of online attractiveness predictors, mostly aimed at western audiences, don't necessarily represent the state-of-the-art in FAP, which seems currently dominated by east Asian research (primarily China), and corresponding east Asian datasets.

Dataset examples from the 2020 paper 'Asian Female Facial Beauty Prediction Using Deep Neural Networks via Transfer Learning and Multi-Channel Feature Fusion'. Source: https://www.semanticscholar.org/paper/Asian-Female-Facial-Beauty-Prediction-Using-Deep-Zhai-Huang/59776a6fb0642de5338a3dd9bac112194906bf30

Broader commercial uses for beauty estimation include online dating apps, and generative AI systems designed to 'touch up' real avatar images of people (since such applications required a quantized standard of beauty as a metric of effectiveness).

Drawing Faces

Attractive individuals continue to be a valuable asset in advertising and influence-building, making the financial incentives in these sectors a clear opportunity for advancing state-of-the-art FAP datasets and frameworks.

For instance, an AI model trained with real-world data to assess and rate facial beauty could potentially identify events or individuals with high potential for advertising impact. This capability would be especially relevant in live video streaming contexts, where metrics such as 'followers' and 'likes' currently serve only as *implicit* indicators of an individual's (or even a facial type's) ability to captivate an audience.

This is a superficial metric, of course, and voice, presentation and viewpoint also play a significant role in audience-gathering. Therefore the curation of FAP datasets requires human oversight, as well as the ability to distinguish facial from 'specious' attractiveness (without which, out-of-domain influencers such as Alex Jones could end up affecting the average FAP curve for a collection designed solely to estimate facial beauty).

LiveBeauty

To address the shortage of FAP datasets, researchers from China are offering the first large-scale FAP dataset, containing 100,000 face images, together with 200,000 human annotations estimating facial beauty.

Samples from the new LiveBeauty dataset. Source: https://arxiv.org/pdf/2501.02509

Entitled *LiveBeauty*, the dataset features 10,000 different identities, all captured from (unspecified) live streaming platforms in March of 2024.

The authors also present FPEM, a novel multi-modal FAP method. FPEM integrates holistic facial prior knowledge and multi-modal aesthetic semantic features via a Personalized Attractiveness Prior Module (PAPM), a Multi-modal Attractiveness Encoder Module (MAEM), and a Cross-Modal Fusion Module (CMFM).

The paper contends that FPEM achieves state-of-the-art performance on the new LiveBeauty dataset, and other FAP datasets. The authors note that the research has potential applications for enhancing video quality, content recommendation, and facial retouching in live streaming.

The authors also promise to make the dataset available 'soon' – though it must be conceded that any licensing restrictions inherent in the source domain seem likely to pass on to the majority of applicable projects that might make use of the work.

The new paper is titled *Facial Attractiveness Prediction in Live Streaming: A New Benchmark and Multi-modal Method*, and comes from ten researchers across the Alibaba Group and Shanghai Jiao Tong University.

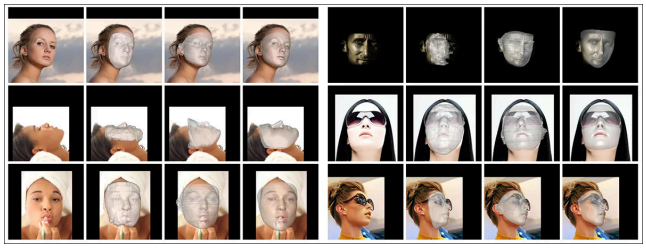
Method and Data

From each 10-hour broadcast from the live streaming platforms, the researchers culled one image per hour for the first three hours. Broadcasts with the highest page views were selected.

The collected data was then subject to several pre-processing stages. The first of these is *face region size measurement*, which uses the 2018 CPU-based FaceBoxes detection model to generate a bounding box around the facial lineaments. The pipeline ensures the bounding box's shorter side exceeds 90 pixels, avoiding small or unclear face regions.

The second step is *blur detection*, which is applied to the face region by using the variance of the Laplacian operator in the height (Y) channel of the facial crop. This variance must be greater than 10, which helps to filter out blurred images.

The third step is *face pose estimation*, which uses the 2021 3DDFA-V2 pose estimation model:



Examples from the 3DDFA-V2 estimation model. Source: https://arxiv.org/pdf/2009.09960

Here the workflow ensures that the pitch angle of the cropped face is no greater than 20 degrees, and the yaw angle no greater than 15 degrees, which excludes faces with extreme poses.

The fourth step is *face proportion assessment*, which also uses the segmentation capabilities of the 3DDFA-V2 model, ensuring that the cropped face region proportion is greater than 60% of the image, excluding images where the face is not prominent. i.e., small in the overall picture.

Finally, the fifth step is *duplicate character removal*, which uses a (unattributed) state-of-the-art face recognition model, for cases where the same identity appears in more than one of the three images collected for a 10-hour video.

Human Evaluation and Annotation

Twenty annotators were recruited, consisting of six males and 14 females, reflecting the demographics of the live platform used*. Faces were displayed on the 6.7-inch screen of an iPhone 14 Pro Max, under consistent laboratory conditions.

Evaluation was split across 200 sessions, each of which employed 50 images. Subjects were asked to rate the facial attractiveness of the samples on a score of 1-5, with a five-minute break enforced between each session, and all subjects participating in all sessions.

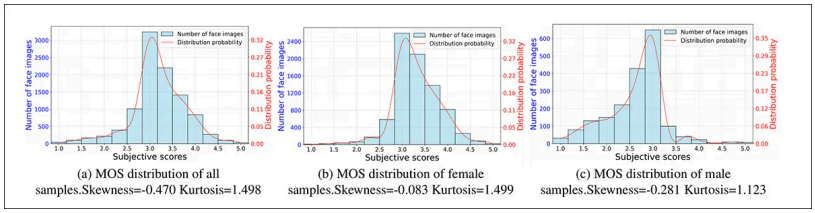
Therefore the entirety of the 10,000 images were evaluated across twenty human subjects, arriving at 200,000 annotations.

Analysis and Pre-Processing

First, subject post-screening was performed using outlier ratio and Spearman's Rank Correlation Coefficient (SROCC). Subjects whose ratings had an SROCC less than 0.75 or an outlier ratio greater than 2% were deemed unreliable and were removed, with 20 subjects finally obtained..

A Mean Opinion Score (MOS) was then computed for each face image, by averaging the scores obtained by the valid subjects. The MOS serves as the ground truth attractiveness label for each image, and the score is calculated by averaging all the individual scores from each valid subject.

Finally, the analysis of the MOS distributions for all samples, as well as for female and male samples, indicated that they exhibited a Gaussian-style shape, which is consistent with real-world facial attractiveness distributions:



(a) MOS distribution of all samples.Skewness=-0.470 Kurtosis=1.498 (b) MOS distribution of female samples.Skewness=-0.083 Kurtosis=1.499 (c) MOS distribution of male samples.Skewness=-0.281 Kurtosis=1.123

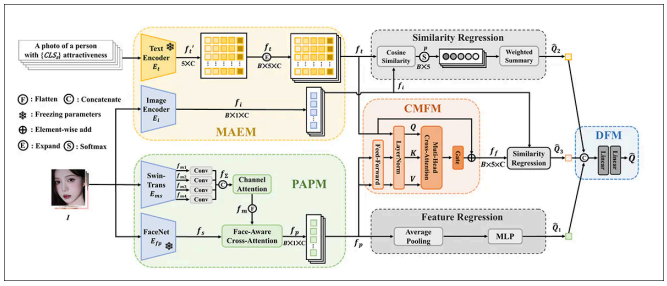
Examples of LiveBeauty MOS distributions.

Most individuals tend to have average facial attractiveness, with fewer individuals at the extremes of very low or very high attractiveness.

Further, analysis of skewness and kurtosis values showed that the distributions were characterized by thin tails and concentrated around the average score, and that *high attractiveness was more prevalent among the female samples* in the collected live streaming videos.

Architecture

A two-stage training strategy was used for the Facial Prior Enhanced Multi-modal model (FPEM) and the Hybrid Fusion Phase in LiveBeauty, split across four modules: a Personalized Attractiveness Prior Module (PAPM), a Multi-modal Attractiveness Encoder Module (MAEM), a Cross-Modal Fusion Module (CMFM) and the a Decision Fusion Module (DFM).



Conceptual schema for LiveBeauty's training pipeline.

The PAPM module takes an image as input and extracts multi-scale visual features using a Swin Transformer, and also extracts face-aware features using a pretrained FaceNet model. These features are then combined using a cross-attention block to create a personalized 'attractiveness' feature.

Also in the Preliminary Training Phase, MAEM uses an image and text descriptions of attractiveness, leveraging CLIP to extract multi-modal aesthetic semantic features.

The templated text descriptions are in the form of *'a photo of a person with {a} attractiveness'* (where *{a}* can be *bad*, *poor*, *fair*, *good* or *perfect*). The process estimates the cosine similarity between textual and visual embeddings to arrive at an attractiveness level probability.

In the Hybrid Fusion Phase, the CMFM refines the textual embeddings using the personalized attractiveness feature generated by the PAPM, thereby generating personalized textual embeddings. It then uses a similarity regression strategy to make a prediction.

Finally, the DFM combines the individual predictions from the PAPM, MAEM, and CMFM to produce a single, final attractiveness score, with a goal of achieving a sturdy consensus

Loss Functions

For loss metrics, the PAPM is trained using an L1 loss, a a measure of the absolute difference between the predicted attractiveness score and the actual (ground truth) attractiveness score.

The MAEM module uses a more complex loss function that combines a scoring loss (LS) with a merged ranking loss (LR). The ranking loss (LR) comprises a fidelity loss (LR1) and a two-direction ranking loss (LR2).

LR1 compares the relative attractiveness of image pairs, while LR2 ensures that the predicted probability distribution of attractiveness levels has a single peak and decreases in both directions. This combined approach aims to optimize both the accurate scoring and the correct ranking of images based on attractiveness.

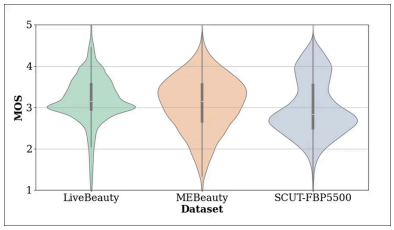
The CMFM and the DFM are trained using a simple L1 loss.

Tests

In tests, the researchers pitted LiveBeauty against nine prior approaches: ComboNet; 2D-FAP; REX-INCEP; CNN-ER (featured in REX-INCEP); MEBeauty; AVA-MLSP; TANet; Dele-Trans; and EAT.

Baseline methods conforming to an Image Aesthetic Assessment (IAA) protocol were also tested. These were ViT-B; ResNeXt-50; and Inception-V3.

Besides LiveBeauty, the other datasets tested were SCUT-FBP5000 and MEBeauty. Below, the MOS distributions of these datasets are compared:



MOS distributions of the benchmark datasets.

Respectively, these guest datasets were split 60%-40% and 80%-20% for training and testing, separately, to maintain consistence with their original protocols. LiveBeauty was split on a 90%-10% basis.

For model initialization in MAEM, VT-B/16 and GPT-2 were used as the image and text encoders, respectively, initialized by settings from CLIP. For PAPM, Swin-T was used as a trainable image encoder, in accordance with SwinFace.

The AdamW optimizer was used, and a learning rate scheduler set with linear warm-up under a cosine annealing scheme. Learning rates differed across training phases, but each had a batch size of 32, for 50 epochs.

Aspects	Methods	LiveBeauty			MEBeauty [6]			SCUT-FBP5500 [5]		
		SROCC↑	PLCC↑	KROCC↑	SROCC↑	PLCC↑	KROCC↑	SROCC↑	PLCC↑	KROCC↑
Baseline	ViT-B [77]	0.897	0.847	0.733	0.620	0.659	0.443	0.869	0.880	0.686
	ResNeXt-50 [78]	0.907	0.879	0.746	0.700	0.743	0.516	0.899	0.911	0.727
	Inception-V3 [79]	0.906	0.880	0.745	0.676	0.727	0.492	0.905	0.918	0.736
IAA	AVA-MLSP [46]	0.833	0.815	0.650	0.602	0.643	0.431	0.764	0.794	0.566
	TANet [47]	0.891	0.858	0.728	0.651	0.690	0.467	0.870	0.882	0.686
	Dele-Trans [80]	0.888	0.841	0.721	0.611	0.651	0.440	0.870	0.860	0.682
	EAT [81]	0.893	0.867	0.728	0.640	0.697	0.462	0.853	0.872	0.665
FAP	ComboNet [31]	0.902	0.871	0.741	0.651	0.692	0.472	0.897	0.907	0.725
	2D-FAP [32]	0.896	0.854	0.733	0.690	0.719	0.506	0.903	0.915	0.734
	REX-INCEP [35]	0.908	0.884	0.750	0.702	0.739	0.514	0.907	0.917	0.739
	CNN-ER [35]	**0.913**	**0.888**	**0.757**	**0.706**	**0.748**	**0.519**	**0.910**	**0.922**	**0.745**
	MEBeauty [6]	0.826	0.796	0.643	0.649	0.677	0.467	0.799	0.802	0.599
	FPEM (Ours)	**0.925**	**0.892**	**0.773**	**0.787**	**0.811**	**0.595**	**0.931**	**0.939**	**0.778**
	Improvement	+1.2%	+0.4%	+1.6%	+8.1%	+6.3%	+7.6%	+2.1%	+1.7%	+3.3%

Results from tests

Results from tests on the three FAP datasets are shown above. Of these results, the paper states:

> *'Our proposed method achieves the first place and surpasses the second place by about 0.012, 0.081, 0.021 in terms of SROCC values on LiveBeauty, MEBeauty and SCUT-FBP5500 respectively, which demonstrates the superiority of our proposed method.*
>
> *'[The] IAA methods are inferior to the FAP methods, which manifests that the generic aesthetic assessment methods overlook the facial features involved in the subjective nature of facial attractiveness, leading to poor performance on FAP tasks.*
>
> *'[The] performance of all methods drops significantly on MEBeauty. This is because the training samples are limited and the faces are ethnically diverse in MEBeauty, indicating that there is a large diversity in facial attractiveness.*
>
> *'All these factors make the prediction of facial attractiveness in MEBeauty more challenging.'*

Ethical Considerations

Research into attractiveness is a potentially divisive pursuit, since in establishing supposedly empirical standards of beauty, such systems will tend to reinforce biases around age, race, and many other sections of computer vision research as it relates to humans.

It could be argued that a FAP system is inherently *predisposed* to reinforce and perpetuate partial and biased perspectives on attractiveness. These judgments may arise from human-led annotations – often conducted on scales too limited for effective domain generalization – or from analyzing attention patterns in online environments like streaming platforms, which are, arguably, far from being meritocratic.

* *The paper refers to the unnamed source domain/s in both the singular and the plural.*

First published Wednesday, January 8, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai