

# A Paradigm for Building Generalized Models of Human Image Perception through Data Fusion

Shaojing Fan<sup>1</sup>, Tian-Tsong Ng<sup>2</sup>, Bryan L. Koenig<sup>3</sup>, Ming Jiang<sup>1</sup>, and Qi Zhao<sup>\*1</sup>

<sup>1</sup>National University of Singapore

<sup>2</sup>Institute for Infocomm Research

<sup>3</sup>Washington University in St. Louis

## Abstract

*In many sub-fields, researchers collect datasets of human ground truth that are used to create a new algorithm. For example, in research on image perception, datasets have been collected for topics such as what makes an image aesthetic or memorable. Despite high costs for human data collection, datasets are infrequently reused beyond their own fields of interest. Moreover, the algorithms built from them are domain-specific (predict a small set of attributes) and usually unconnected to one another. In this paper, we present a paradigm for building generalized and expandable models of human image perception. First, we fuse multiple fragmented and partially-overlapping datasets through data imputation. We then create a theoretically-structured statistical model of human image perception that is fit to the fused datasets. The resulting model has many advantages. (1) It is generalized, going beyond the content of the constituent datasets, and can be easily expanded by fusing additional datasets. (2) It provides a new ontology usable as a network to expand human data in a cost-effective way. (3) It can guide the design of a generalized computational algorithm for multi-dimensional visual perception. Indeed, experimental results show that a model-based algorithm outperforms state-of-the-art methods on predicting visual sentiment, visual realism and interestingness. Our paradigm can be used in various visual tasks (e.g., video summarization).*

## 1. Introduction

When it comes to understanding visual perception of images, studying human judgment remains the leading technique [1]. Scientists have studied various high-level image properties that impact human behavior [2, 3, 4, 5, 6, 7, 8]. Such studies are done with psychophysical experiments in labs or online surveys via crowdsourcing platforms such

as Amazon Mechanical Turk (AMT). Consequently, each study creates an isolated set of data mostly used for a specific research purpose, often containing strong built-in bias [9]. Furthermore, collecting human data is costly and time-consuming. The average pay to AMT workers is \$2-3 per hour, and doing 40 ratings takes about 5 minutes, so to have 10 workers rate each of 10K images for merely 40 attributes would cost around \$21K [10].

We believe that the reuse of existing datasets can provide novel insights. Although various datasets were created with different aims, many are manifestations of a shared underlying human psychological process [11]. We focus on perception of visual images, which we hypothesize to have few dimensions [11, 12]. Indeed, many of the previously studied visual perceptual attributes are interrelated (e.g., image naturalness is related to interestingness [13]). Based on the above understandings, we propose to reuse and fuse existing isolated datasets to statistically and computationally model visual perception (see Fig. 1 for an overview). Our contributions are as follows.

1. *We report a method to create “big human data” from smaller datasets.* We show how to aggregate heterogeneous and fragmented human datasets that partially overlap in content. The non-overlapping parts of dataset are effectively missing data, whose values we infer using multiple imputation to get a “bigger” dataset. Notably, the fused dataset outperforms the constituent datasets when predicting image perceptions (e.g., visual sentiment, visual realism, and interestingness) across more diverse image semantics. Thus, through data fusion we get extra performance for free. Our method provides another path in the journey of achieving *big human data* – the fused dataset includes all images and all attributes from the partially overlapping constituent datasets.
2. *We propose a paradigm for iteratively increasing the generality of a model of human image perception.* Based on structure-revealing data fusion and statistical modeling, we

\*Corresponding author.

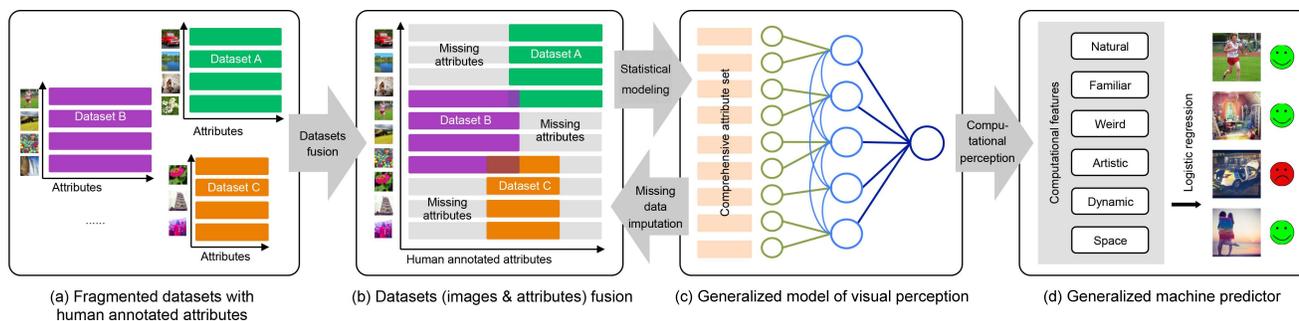


Figure 1: Overview of our approach. First, we fused partially overlapping datasets to form a larger dataset (a, b). Second, we built a generalized and expandable human perception model through structural-revealing statistical modeling, which also informs data fusion (c). Finally, we built a generalized computational algorithm guided by the model, and test its performance by comparing their predictions of multiple image perceptions with alternatives (d).

build a multi-layered model of human image perception of digital images. The model incorporates commonly studied image properties (*e.g.*, aesthetics, interestingness) with properties seldom studied for digital images, such as familiarity and scene dynamics (Fig. 3). Unlike computer paradigms, our theory-based model characterizes human perception in a more understandable and concise way while reducing the danger of overfitting. It provides an improved ontology that can reduce the amount of human data collection needed and therefore associated costs. Critically, we show that it can guide the design of a generalized computational predictor, which is able to predict multiple high level image attributes simultaneously.

## 2. Related work

**Predicting high-level image properties:** A common approach is to link lower-level image attributes with higher-level properties, such as aesthetics [14, 15], interestingness [2], memorability [16], visual realism [3, 8], and emotions [17]. Most relevant is [4], which proposed a visual sentiment ontology together with a set of sentiment descriptors. Although these studies generated algorithms with considerable prediction performance on specific attributes, few insights were provided to explain why the algorithms work. Furthermore, most of the studies are isolated. Despite the exploration in [16] on correlations between memorability and other image properties, there is no clear framework that explains the correlational structure of various perceptions. Instead of isolated goal-specific tasks, we propose a comprehensive ontology that can guide the design for predicting various human perceptions based on an intuitively understandable theory-based model of human perception.

**Visual perception datasets:** Two datasets with extensive human annotations are the Visual Realism Dataset [3] and the Memorability Dataset [5] (See Sec. 3.1 for details). One related dataset consists of affective images datasets [17]; another is the Sentiment Dataset [4]. There are other datasets

built for specific goals, such as aesthetics [14] and interestingness [2]. However, they either have a single rating or binary labels only, and thus are insufficient for visual perception modeling. These datasets are much smaller than popular image datasets (*e.g.*, ImageNet [18]) due to high costs of human data collection. Our work fuses some of these datasets into larger sets, to get a more general understanding of human image perception.

**Visual perception structure:** The human nervous system processes visual information both hierarchically and in parallel [19, 20, 21]. The levels of this processing can be classified as early, intermediate, and late vision [22, 23]. The basic features of the image itself, like orientation and color, are processed during the early vision stage in the occipital lobe’s visual processing areas [24, 25]. Intermediate vision segregates visual elements to form coherent regions in what would otherwise be a chaotic and overwhelming sensory array [26]. Late vision selects which of these coherent regions to scrutinize and evokes emotions and memories from which objects are recognized and meanings attached [20]. Some studies suggest visual perception is low dimensional [12, 11].

**Data fusion and inference:** In many disciplines, from computer vision to neuroscience, data from multiple sources are acquired and jointly modeled for enhanced knowledge discovery [27], but joining data often results in missing data. A visual synsets was proposed in [28] to infer missing attributes through Linear SVM prediction. In psychology and sociology, multiple imputation (MI) is a common method for handling missing data [29]. It estimates multiple values for missing data to account for the uncertainty surrounding missing values, thus providing unbiased estimates and more validity to missing data [30]. We believe our work is the first to apply MI to visual perception for computer vision applications. Our MI was guided by comprehensive data exploration and theory-based understanding of the underlying structure of human perception.

Table 1: Human annotated attributes from two datasets, used for perception modeling. Attributes that are similar across datasets share the same non-black color.

Visual Realism Dataset [3]	Memorability Dataset [5]
<b>Spatial layout:</b> Clean scene? Close-range vs. Distant-view; Neat Space?*†; Common perspective?* <b>Empty space vs. Full space</b>	Enclosed space vs. Open space*; Perspective view vs. Flat view; Mirror symmetry vs. No mirror symmetry; <b>Empty space vs. Cluttered space</b>
<b>Aesthetics and presentation:</b> Sharp vs. Blurry*; Expert photography? Lighting effect natural*? Color appearance natural*? Colors go well together? Colorful*? <b>High quality vs. Low quality; Attractive to you?</b>	Post-card like? Is aesthetic? Pleasant vs. Unpleasant scene* ; Boring vs. Striking colors; Unusual or strange vs. Routine or mundane; <b>High quality (expert photography) vs. Poor quality photo; Attractive vs. Dull photo</b>
<b>Feelings:</b> Familiar with the scene? Familiar with the objects?* <b>Unusual or Strange?; Mysterious?; Makes you happy?; Makes you sad?†; Exciting?</b>	Frightening* ? Arousing*†? Funny? Engaging? Peaceful? Interesting*? Striking? <b>Strange? Mysterious? Makes you happy? Makes you sad?† Exciting?</b>
<b>Semantics:</b> Contain fine details*? Object appearance natural*? Naturally-occurring objects combinations?* <b>Dynamic or energetic scene? Is there a storyline?</b>	Action going on? Something moving in scene? About to happen? Lot going on? Have a lot to say; Length of description; <b>Dynamic scene? Static scene?* Picture tells a story?</b>

\* Attributes for imputation in Sec. 4. † Selected as auxiliary variables [30] in data fusion in Sec. 4.

### 3. Empirical modeling on separate datasets

Our goal was to fuse multiple datasets to get a broader picture of human perception. To prepare for data fusion, we first separately explored two datasets with human annotations to get partial pictures of human image perception. We built human perception models from each dataset through statistical modeling to identify common and distinctive factors across datasets, as well as their latent structures. Such structures also guide data fusion in Sec. 4.1.

#### 3.1. Datasets with human perception

We used the following two datasets in empirical modeling, and the data fusion in Sec. 4.

**Visual Realism Dataset** [3] was collected to study image visual realism. It includes 2520 images of diverse scenes, half of which are computer generated graphics and the rest are photographs. Each image has 40 human annotations (hereafter referred to as *attributes*).

**Memorability Dataset** [5] was collected to study image memorability. It includes 2222 images of general scenes. Each image has 127 human-annotated attributes.

For both datasets, attributes were collected using AMT. Those used for perception modeling are shown in Table 1. The complete lists are shown in the supplementary material.

#### 3.2. Model Structure

We designed our model structure based on previous findings in psychology, neuroscience, and linguistics. Human visual perception occurs in a hierarchical fashion [19]. Correspondingly, we hypothesized that the human perception model is hierarchically structured with three layers. The first layer includes “basic perceptual features” of the image itself, such as texture and shape, that are analogous to those processed in the visual cortex [31]. The second layer integrates the lower-level information. For example,

determining whether the scene in the image is familiar. The third layer corresponds to effective (liking) reactions to images [12]. Our model emphasizes the structure of the upper two layers, which are higher-level human perception. For the lower level we merely used attributes from the datasets to feed into the second, inferred layer. The structure is also inspired from WordNet [32], in which the vocabulary of a language is mapped to a set of concepts. In the next few sections, we conducted statistical modeling to fit the data to the three-level model and test the goodness of fit.

#### 3.3. Statistical modeling

Based on the hierarchical structure, we created statistical models of human perception separately for each dataset. We first conducted exploratory factor analysis to determine the number of latent variables (factors) present in the human data and identify a compact set of attributes that underlie latent variables. We then performed confirmatory factor analysis to test relations between latent and observed variables and assess the reliability and validity of measures. Finally, path analysis was conducted to determine standardized regression weights ( $\gamma$ ) and correlations ( $\phi$ ) among latent variables [33]

**Attribute pruning:** Our aim was to build a model that reflects human perception, so we excluded non-subjective attributes, such as “sky present?”. We also excluded attributes that were unrelated to perception or likely biased by personal experiences such as “recognize place?”. The final number of attributes used to create human perception models for the two datasets were 25 and 32 respectively (see Table 1).

**Exploratory and confirmatory factor analysis:** To measure how observed variables (attributes) are related to latent variables (higher-level perceptions and reactions), we performed exploratory factor analysis (EFA) followed by confirmatory factor analysis (CFA) [33]. Factor analysis (FA) is similar to principal components analysis (PCA); both

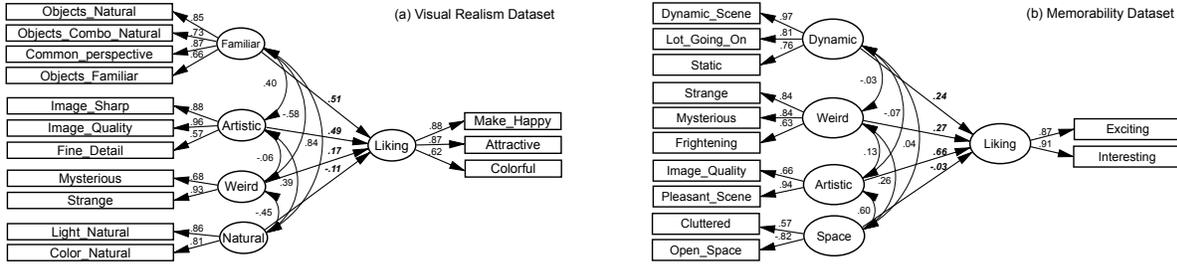


Figure 2: Human perception models built on two separate datasets. Standardized regression weights ( $\gamma$ ) are in bold italic font. Although collected for different research purposes, the two datasets shared two latent factors (*weird* and *artistic*), indicating that common perceptual processes were measured in different datasets. The 25 and 32 attributes in each dataset were encapsulated by 5 dimensions respectively (four layer-1 factors in the lower layer and one layer-2 factor in the upper layer, *liking*), suggesting the multidimensionality and low dimensionality of visual perception.

are variable reduction techniques. However, in PCA, the components are orthogonal linear combinations that maximize the total variance, whereas in FA, the factors are linear combinations that maximize the shared portion of the variance underlying “latent constructs”. We applied maximum likelihood in EFA with oblique transformation (where the new axes can be non-orthogonal, thus allowing factors to correlate [34]), as we hypothesized that visual perception is multidimensional. A CFA [33] performed after EFA tested how well the measured variables represent the number of constructs identified in FA. Attributes with poor loadings or fits were eliminated. Detailed results are reported in supplementary material.

**Path analysis:** Guided by the designed model structure in Sec. 3.2, we divided the latent factors from previous CFA into two layers (Fig. 2). The final model was created through a path analysis predicting the *liking* latent construct from the lower-level perception latent constructs. Path analysis is a straightforward extension of multiple regression, which aims to provide concurrently estimated magnitudes of hypothesized causal connections between sets of variables [35]. Readers can refer to [36] for its detailed formulae and derivation. The final models for each dataset are shown in Fig. 2. For all models in the paper, detailed estimates are reported in supplementary material.

We applied two common metrics to measure the fitness of the model to the data. The first is Comparative Fit Index (*CFI*), which compares a chi-square for the fit of a target model to the fit of an independent model—one in which the variables are uncorrelated. Higher *CFI* indicates greater extent to which the model of interest is better than the independence model. Values that approach .90 indicate acceptable fit [33]. Another model fit metric is Root Mean Square Error of Approximation (*RMSEA*), which estimates the amount of error of approximation per model degree of freedom and takes sample size into account. Smaller *RMSEA* values suggest better model fit. A value of .10 or less is indicative of acceptable model fit [33]. Our models had acceptable fit,

$CFIs \geq .93, RMSEA \leq .092$ .

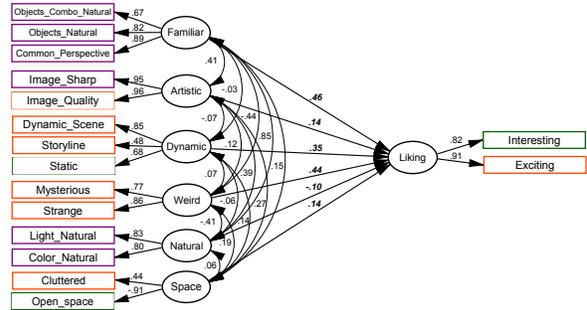


Figure 3: Human perception model (fused model 1) based on the fusion of two datasets. Those with purple borders are from Visual Realism Dataset, those with green borders are from Memorability Dataset. Similar attributes from both datasets have orange borders.

## 4. Data fusion

To develop a more general human perception model, we fused the two datasets. Data fusion was enabled by shared perceptions across the datasets.

### 4.1. Inferring missing data

Some attributes are in both datasets, but the attributes in only one dataset result in missing values on the images from the other dataset. One common approach for inferring missing values is to replace them with the mean, the median or the modal value [37]. However, we hypothesized that human perception has a low-dimensional structure, and we expected components to be highly correlated, so we chose MI which utilizes all available data to preserve sample size and statistical power.

We started with attributes pruning: we collapsed 10 similar attributes from the datasets (color-paired in Table 1). We further included 16 additional attributes that showed up in separate models (4 of which were used as *auxiliary vari-*

ables<sup>1</sup> [30]). We did not include all attributes as doing so would introduce excessive missing data [30]. Attributes for imputation are marked with \* in Table 1.

We performed MI using the Markov Chain Monte Carlo method with a noninformative prior and a single chain [38]. The multivariate normal model was used for all imputations. Convergence and independence of imputed values were assessed with time series and autocorrelation plots of means and variances [29]. Specifically, we denote the variables with missing values for observation  $i$  by  $Y_{mis}^{(i)}$  and the variables with observed values by  $Y_{obs}^{(i)}$ , the iteration goes as follows: with a current parameter estimate  $\theta^{(t)}$  at the  $t$ th iteration, a first step will draw  $Y_{mis}^{(t+1)}$  from  $p(Y_{mis}^{(t+1)}|Y_{obs}^{(t)}, \theta^{(t)})$  and a second step draws  $\theta^{(t+1)}$  from  $p(\theta|Y_{obs}^{(t)}, Y_{mis}^{(t+1)})$ . This creates a Markov chain

$$(Y_{mis}^{(1)}, \theta^{(1)}), (Y_{mis}^{(2)}, \theta^{(2)}), \dots \quad (1)$$

which converges in distribution to  $p(Y_{mis}, \theta|Y_{obs})$ .

We ran MI five times to minimize standard errors [39], resulting in five datasets with both observed and imputed values. We refer to those datasets as *multiple fused datasets*.

## 4.2. Empirical modeling

The model structure and modeling steps were the same as in Sec. 3.3. Overall estimates were the means of each fused dataset’s estimates [29, 30]. This averaging was done for all subsequently reported results of imputed datasets. The resulting perception model (hereafter, *fused model 1*) is shown in Fig. 3. It had acceptable fit,  $CFI = .92$ ,  $RMSEA = .089$ .

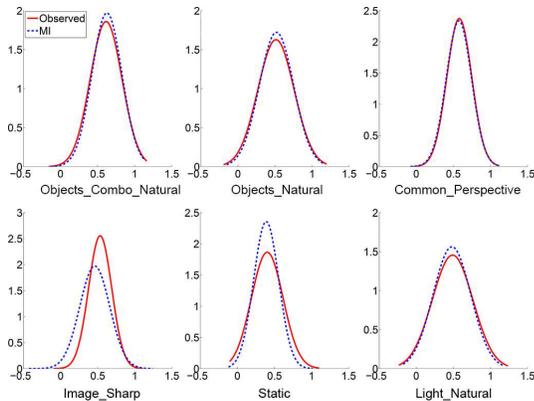


Figure 4: Frequency distribution (by normal fitting) of observed and imputed values of the first six attributes in fused model 1. For a complete list see supplementary material.

<sup>1</sup>Auxiliary variables are observed variables that correlate with the missing values but are not part of any factors used in the model itself. They can improve missing data imputation [30]. Their selection was based on prior modeling of separate datasets (see Sec. 3.3).

## 4.3. MI evaluation

The validity of MI was evaluated four ways. First we compared frequency distributions of the observed attribute values to those generated by MI. This is the most common evaluation approach [30]. The observed and imputed data had similar distributions (Fig. 4), suggesting consistency from imputation. Second, we compared estimates produced by the MI model with those produced by models using only the observed variables. The estimates were consistent (see supplementary material), indicating that MI retained the model structure. We further compared MI by replacing missing values by means. Although they produced similar predictions for the layer-2 factor, the attributes from MI are more normally distributed thus better for path analysis [35]. Finally, we compared cross-factor correlations for factor combinations present in only one of the separate models with their correlations in fused model 1. Such correlations were highly consistent (see Fig. 3 and 2). The coefficients of different factors to *liking* were also consistent in terms of strength and sign. This indicates that the fused model is largely similar to the constituent models, except it is better than either because it includes attributes from both. In other words, *data fusion allows more latent factors to be included, thus generating a visual perception model more general than its constituents*.

## 4.4. Discussion

**Generality and expandability:** Fused model 1 and the separate models shared latent structure comprised of a small number of common factors, which indicates generality and low dimensionality of the model. Similarly, research has suggested that human emotions [4] and brain structures for perceiving visual stimuli [12] are low dimensional. Notably, our model can be extended easily by adding new nodes or latent factors whenever new datasets are available.

**Digging deeper into the model:** We compare our findings with three commonly studied human perceptions:

1. *Liking*: In psychology, several factors have been found to influence interpersonal attraction, namely proximity, familiarity, similarity, and physical attractiveness [40, 41]. This is reminiscent of our model of visual perception, in which the factors *familiarity*, *naturalness*, and *artistic* are analogous to familiarity, similarity, and physical attractiveness, respectively. The strong weights from *familiarity* to *liking* is also consistent with previous psychological studies [42, 43]. Interestingly, the weight of *familiarity* to *liking* is significantly stronger than that of *artistic*, whereas it is comparable with that of *weird* (Fig. 3). This suggests that visual preference of an image is more influenced by the specific feelings aroused in humans rather than its aesthetic value per se. What’s more, *liking* has a strong loading node, *exciting*, which accords with previous psychological findings that exciting is a combination of pleasure and arousal [44].

2. *Artistic*: Image sharpness and image quality strongly loaded on *artistic*. This provides support to previous studies which used similar features, such as image contrast and edge distribution in aesthetics evaluation [14, 15]. *Artistic* strongly correlated with *familiarity* and *naturalness* (Fig. 3). This again supports [15] in which familiarity was used as one feature dimension for image aesthetics.

3. *Naturalness*: Colorfulness, image sharpness, reproduction of shadow detail, and absence of washed-out appearance were found to be important factors for naturalness [45]. This is in line with our model, in which *naturalness* is strongly loaded by *lighting naturalness* and *color naturalness*. Our model suggests that *naturalness* strongly correlates with *familiarity* ( $\phi = .85$ ). Thus the nodes of *familiarity* (natural objects combination and appearance, common perspective) also contribute to *naturalness*. Similarly, [13] suggests that image semantics such as object combination and viewing perspective are crucial for naturalness.

In summary, our model summarizes and visualizes multi-dimensional human perception in a concise and understandable way. It provides a new ontology for studies that predict high-level image properties, while also corroborating previous findings in visual perception.

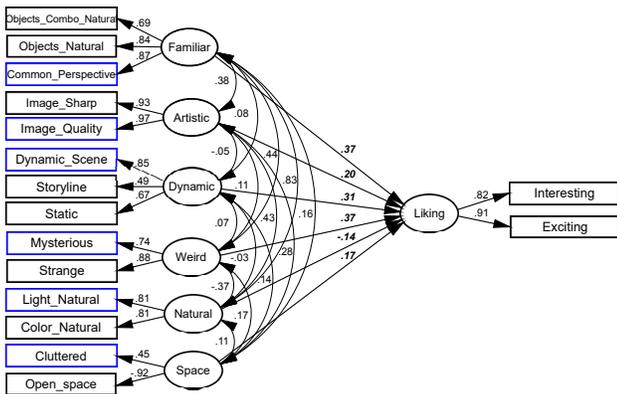


Figure 5: Human perceptual model (fused model 2) based on the fusion of three datasets. In Sentiment Dataset, attributes with blue borders were collected from survey on AMT, other attributes were derived through imputation.

## 5. Predicting visual sentiment, visual realism, and interestingness

In this section, we demonstrate how to build a generalized computer algorithm based on the model and data fusion. We incorporate a new dataset into a fused dataset for the purpose of predicting multiple perceptions: visual sentiment, visual realism, and interestingness. We show that our perception model can be used to guide both survey design and building machine algorithms for predicting various human perceptions.

## 5.1. Data fusion

**Sentiment Dataset:** We used the Sentiment Dataset in all our experiments on perception prediction. It includes 603 tweets with photos and was originally collected to evaluate the performance of an automatic sentiment prediction method. Ground truth of binary sentiment values were obtained by human annotation using AMT [4].

**Model structure:** To leverage of our fused model, we mapped sentiment onto *liking*. Two observations validate this mapping. First, the *makes happy* attribute is central to positive sentiment [4]. Second, the average Spearman’s rank correlation between *exciting* and *interesting* (which strongly loaded on *liking*) with *makes happy* on Visual Realism and Memorability datasets was .71, suggesting that exciting and interesting were highly indicative of positive sentiment. So we regarded the *liking* factor as positive sentiment and reused our previous model structure.

**Human survey:** The Sentiment Dataset had only binary labels so we could not fuse it into our perception model. Therefore, we conducted a survey on AMT, to collect human annotations for the dataset. Our perception model guided the design of the survey: we wanted to restrict our survey questions to the observed attributes in our model because these attributes enable us to predict perceptual factors other than sentiment. We randomly selected a child node of each layer-1 factor in Fig. 3, resulting in a six-question survey which had no direct connection to *liking* in fused model 1 (see Fig. 5 for selected questions). We will show that using multiple imputation and the empirical models enables these 6 questions not only to predict sentiment, but also to predict multiple perceptual factors in our network.

**Perception model based on fusion of three datasets:** The fusion and modeling procedures were the same as in Sec. 4. The result, *fused model 2* (Fig. 5) has acceptable fit,  $CFI = .93$ ,  $RMSEA = .089$ . Its estimates are highly consistent with fused model 1 (Fig. 3).

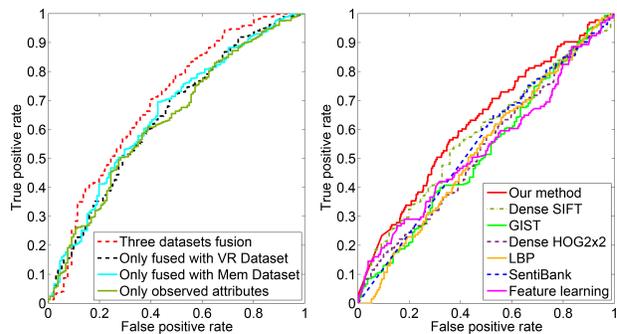


Figure 6: Sentiment prediction using human attribute ratings (left) and machine features (right) as inputs.

## 5.2. Computational algorithm design

To build a generalized computational predictor, we computerized the empirical model by designing computational features to predict its layer-1 factors (see Fig 5 for the factors). We modeled *natural* and *familiar* (which were highly correlated,  $\phi = .83$ ) using [46], in which scale invariant natural image patterns were modeled through high-order image patch statistics. Primarily meant for image quality assessment, we used it here to quantify naturalness. We used Ke’s method on aesthetics rating [14] for the *artistic* factor, which includes perceptual motivated features such as image composition, blur, and contrast. We applied GIST descriptors [47] to model *space*. They provide a statistical summary of the spatial layout properties (e.g. openness, expansion) of the scene. Finally, we detected global outliers as for *weird* factor by applying the Local Outlier Factor (LOF) algorithm [48] to global image descriptors (GIST and SIFT [49]).

Table 2 shows the list of our features. The detailed design is reported in supplementary material. We did not find niche features for the *dynamics* factor, as it is highly abstract. However, scene dynamics is somewhat captured by the Spatial Envelope Model in the GIST descriptor, and the blur component in artistic features. We then combined the features in later fusion according to the weights of the links towards *liking* in fused model 2 (Table 2).

Table 2: Features for computational prediction.

Perception factor	Weights	Computer feature	Dim
<i>Natural, Familiar</i>	.51 <sup>1</sup>	Natural image statistics [46]	72
<i>Artistic</i>	.20	Ke’s method [14]	12
<i>Space</i>	.17	GIST [47]	512
<i>Weird</i>	.37	LOF [48]	3

<sup>1</sup> Sum of the absolute weights of *natural* and *familiar*.

## 5.3. Predicting sentiment with human data

First of all, we hoped to test the effects of data fusion. We used human ratings of attributes to predict visual sentiment. We did three experiments. In experiment 1, we used all attributes—observed and imputed—in fused model 2 (14 nodes on the left in Fig. 5). In experiment 2, we only used observed attributes—the 6 survey questions. In experiment 3, to see if fusing more datasets made a difference, we built the perception models for fusions of only pairs of datasets (i.e., Sentiment and Visual Realism Dataset, Sentiment and Memorability Dataset). The models’ structures were the same as Fig. 5 except that their coefficients differed. In each experiment, human attribute ratings were used to train SVM [50] for binary sentiment classification. We used grid search to select cost, RBF kernel parameter  $\gamma$ , and  $\epsilon$  hyperparameters. We split the data into 80% as a training set and 20% as a test set. We used area under ROC curve (AUC) as an evaluation metric. As shown in Fig. 6a, using all the

attributes based on the fusion of three datasets produced the best performance, suggesting the advantage of data fusion.

## 5.4. Predicting sentiment with machine features

We tested our model-based algorithm on binary classification of sentiment with the same SVM settings as Sec. 5.3. Late fusion was performed as weighted average of the regression outputs for respective perceptual factors. We compared the result with four baselines, namely dense SIFT [51, 49], GIST [47], dense HOG2x2 [52, 53], and LBP [54]. Our method outperformed all the baselines (Fig. 6b).

We further compared our algorithm with two state-of-the-art algorithms: SentiBank descriptor [4] and unsupervised feature learning. First, we replicated the experiment in [4] using the same set of features on the same Twitter Dataset. Second, we constructed an unsupervised feature learning framework with a single-layer triangular K-means encoding [55] on image patches preprocessed by local intensity and contrast normalization, as well as whitening. We scanned each image with 16-by-16 pixel receptive field and 1 pixel stride, before mapping the preprocessed image patches to 512-dimensional feature vectors. Again our method excelled (Fig. 6b).

## 5.5. Predicting visual realism and interestingness

Being generalized and expandable, our model can be applied to predict other perceptual factors such as visual realism (the degree an image appears to be a photo rather than computer generated [3]) and interestingness, by modifying the layer-2 factor to the corresponding perception (see supplementary material for modified models). We used the same human attributes and computer algorithm as Sec. 5.3 and 5.2, respectively. We performed both support vector regression and binary support vector classification, using the same setting as Sec. 5.3. Features were fused through weighted kernel sum based on their corresponding loadings to layer-2 factors. Ground-truth ratings for visual realism and interestingness were collected using AMT. For binary classification, images with ground-truth realism score  $> .50$ , and interestingness score  $> .75$  ([2]) were separately taken as positive, negative otherwise.

As shown in Table 3, our complete attributes set from data fusion (1st row) performed the best among all evaluation metrics. Our computational features motivated by the perception model (3rd row) produced consistently better results than the state-of-the-art algorithms (4th row). Sample images with computational predictions are shown in Fig. 7.

## 5.6. Discussion

**Power of data fusion:** Using all data in fused datasets was consistently better than using only observed data, and the performance based on the fusion of three datasets was better than that based on fusions of two datasets (Fig. 6a, Table 3).

Table 3: Experiment results for visual realism and interestingness.  $\rho_v$  and  $A_v$  are the Spearman’s rank correlation and AUC for visual realism, respectively.  $\rho_i$  and  $A_i$  are for interestingness.

Feature type	Regression		Classification	
	$\rho_v$	$\rho_i$	$A_v$	$A_i$
All attributes	.70	.71	.90	.75
Only observed attributes	.69	.70	.84	.64
Our method	.54	.31	.82	.71
[3], [2] <sup>1</sup>	.51	.27	.77	.59

<sup>1</sup> In the same row,  $\rho_v$  and  $A_v$  are based on the algorithm in [3],  $\rho_i$  and  $A_i$  are based on the algorithm in [2].

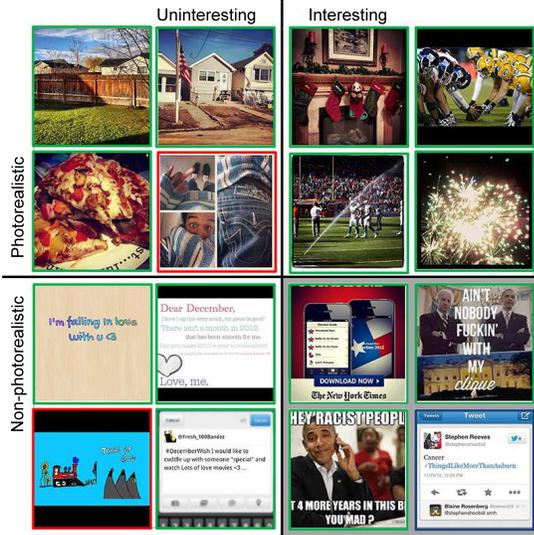


Figure 7: Sample images with computational predictions. Images with green border have positive sentiment, red border for negative sentiment. The bottom right image with blue border was misclassified as positive sentiment (due to our algorithm’s inability to understand text). The four images in lower right quadrant (with grey background) are misclassified as being interesting (partly due to the high threshold of ratings (.75) for being interesting). The rest are correctly classified on interestingness, visual realism, and sentiment.

This suggests that *data imputation boosts performance*. Recall that we only collected 6 attributes, each from one latent factor, and after imputation we had 14 attributes which predicted better than the 6 attributes. In other words, we got the extra performance for free out of a *representative* subset of human annotation through data fusion.

Does the performance gain come from information gain due to data imputation? The answer is probably *no* according to the data processing inequality of information theory [56]. The information gain most likely comes from the expansion of observables due to the data fusion per se, and data imputation makes the usage of off-the-shelf classifiers possible. Nevertheless, the inferred perception model provides

effective guidance for human data collection, if needed, to achieve true information gain.

**Power of the perception model:** We have already shown how the perception model guided the human survey design and data fusion. More importantly, the model is able to guide the design of a generalized computer algorithm for various image perceptions. Our algorithm was consistently better than its alternatives in sentiment prediction (Fig. 6b) and visual realism and interestingness estimation (Table 3). Notably, our algorithm outperformed the original algorithms built from the constituent datasets. This suggests that *our perception model can guide the design of a generalized computational predictor, which produces more human-consistent results*. Note that for sentiment prediction, the closest two methods were dense SIFT and SentiBank. However, dense SIFT was sampled at multiple scales and learned a dictionary of codewords, which finally generated a feature of 3000 dims. SentiBank detectors were trained on about 500k Flickr images [4]. In contrast, our algorithm is based on a fused dataset of only 4603 images and sparse features (599 dim). In short, our approach based on the structure of human perception beats both a highly data-driven and a dense pooling approach that lack of a visual perception model.

**Holistic perception:** By simply changing the weights of the features, the same algorithm can be used to simultaneously predict visual sentiment, interestingness and realism while holding a consistent correlation among these properties. This provides a holistic computation in which different dimensions of human perception share a consistent latent structure. More attributes can be computed by expanding our perception model (*e.g.*, persuasiveness [6], trustworthiness [57]).

## 6. Conclusion and future work

In this paper, we propose a paradigm for building generalized and expandable models of human image perception. Our approach sheds new light on creating “big human data”. The resulting model provides a novel ontology for design of both human surveys and generalized computer algorithm design for multiple visual perceptions. One limitation of our approach is that we used only linear modeling but human perception might be non-linear. Therefore, direct prediction from MI and SEM are inferior to non-linear SVM (see supplementary materials for detailed experiments). Our future work will include non-linear models such as Isomap [58], hoping to get better prediction results from the model itself and boost performance through fusing larger scale datasets.

## 7. Acknowledgement

We would like to thank Robert Kirkpatrick and Michael Neale for helpful discussions on statistical modeling. This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centre in Singapore Funding Initiative.

## References

- [1] Alex Peysakhovich and Stephens-Davidowitz. How not to drown in numbers. [http://www.nytimes.com/2015/05/03/opinion/sunday/how-not-to-drown-in-numbers.html?\\_r=0](http://www.nytimes.com/2015/05/03/opinion/sunday/how-not-to-drown-in-numbers.html?_r=0).
- [2] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. The interestingness of images. In *ICCV*, pages 1633–1640. IEEE, 2013.
- [3] Shaojing Fan, Tian-Tsong Ng, Jonathan S Herberg, Bryan L Koenig, Cheston Y-C Tan, and Rangding Wang. An automated estimator of image visual realism based on human cognition. In *CVPR*, pages 4201–4208. IEEE, 2014.
- [4] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM Multimedia*, pages 223–232, 2013.
- [5] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. Understanding the intrinsic memorability of images. In *NIPS*, pages 2429–2437, 2011.
- [6] Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. Visual persuasion: Inferring communicative intents of images. In *CVPR*, 2014.
- [7] Arturo Deza and Devi Parikh. Understanding image virality. *arXiv preprint arXiv:1503.02318*, 2015.
- [8] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Learning a discriminative model for the perception of realism in composite images. *arXiv preprint arXiv:1510.00477*, 2015.
- [9] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *ECCV*, pages 158–171. Springer, 2012.
- [10] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.
- [11] Bolei Zhou and Liqing Zhang. A hierarchical model for visual perception. In *Advances in Cognitive Neurodynamics (II)*, pages 607–613. Springer, 2011.
- [12] Bilal H Fadlallah, Sohan Seth, Andreas Keil, and José C Príncipe. Analyzing dependence structure of the human brain in response to visual stimuli. In *ICASSP*, pages 745–748. IEEE, 2012.
- [13] Raisa Halonen, Stina Westman, and Pirkko Oittinen. Naturalness and interestingness of test images for visual quality evaluation. In *IS&T/SPIE Electronic Imaging*, pages 78670Z–78670Z, 2011.
- [14] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *CVPR*, volume 1, pages 419–426. IEEE, 2006.
- [15] Ritendra Datta, Jia Li, and James Ze Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *ICIP*, pages 105–108. IEEE, 2008.
- [16] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *ICCV*, 2015.
- [17] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM Multimedia*, pages 83–92. ACM, 2010.
- [18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.
- [19] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.
- [20] Martha J Farah. *The cognitive neuroscience of vision*. Blackwell Publishing, 2000.
- [21] Hillel Adesnik, William Bruns, Hiroki Taniguchi, Z Josh Huang, and Massimo Scanziani. A neural circuit for spatial summation in visual cortex. *Nature*, 490(7419):226–231, 2012.
- [22] David H Warren. Early vs. late vision: The role of early vision in spatial reference systems. *New Outlook for the Blind*, 1974.
- [23] E Goldstein. *Sensation and perception*. Cengage Learning, 2013.
- [24] William E Vinje and Jack L Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.
- [25] Tim Gollisch and Markus Meister. Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron*, 65(2):150–164, 2010.
- [26] Kenneth J Hayworth and Irving Biederman. Neural evidence for intermediate representations in object recognition. *Vision research*, 46(23):4024–4031, 2006.
- [27] Bahador Khaleghi, Alaa Khamis, Fakhreddine O Karray, and Saiedeh N Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44, 2013.
- [28] David Tsai, Yushi Jing, Yi Liu, Henry Rowley, Sergey Ioffe, James M Rehg, et al. Large-scale image annotation using visual synset. In *ICCV*, pages 611–618. IEEE, 2011.
- [29] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
- [30] Linda M Collins, Joseph L Schafer, and Chi-Ming Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, 6(4):330, 2001.
- [31] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.
- [32] Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.
- [33] Rex.B. Kline. *Principles and Practice of Structural Equation Modeling*. Guilford Press, 2011.
- [34] Jae-On Kim and Charles W Mueller. *Introduction to factor analysis: What it is and how to do it*. Number 13. Sage, 1978.
- [35] John C Loehlin. *Latent variable models: An introduction to factor, path, and structural analysis*. Lawrence Erlbaum Associates Publishers, 1998.

- [36] Sewall Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934.
- [37] Imputing missing values before building an estimator. [http://scikit-learn.org/stable/auto\\_examples/missing\\_values.html#example-missing-values-py](http://scikit-learn.org/stable/auto_examples/missing_values.html#example-missing-values-py).
- [38] Yang C Yuan. Multiple imputation for missing data: Concepts and new development (version 9.0). *SAS Institute Inc, Rockville, MD*, 49, 2010.
- [39] Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996.
- [40] Elaine Hatfield and Richard L Rapson. Similarity and attraction in close relationships. *Communications Monographs*, 59(2):209–212, 1992.
- [41] Ellen Berscheid and Harry T Reis. Attraction and close relationships. *The handbook of social psychology*, 2:193–281, 1998.
- [42] John D Balling and John H Falk. Development of visual preference for natural environments. *Environment and Behavior*, 14(1):5–28, 1982.
- [43] Moshe Bar and Mital Neta. Humans prefer curved visual objects. *Psychological science*, 2006.
- [44] James A Russel and Geraldine Pratt. A description of affective quality attributed to environment. *Journal of Personality and Social Psychology*, 38(2):311–322, 1980.
- [45] Seo Young Choi, M Luo, Michael Pointer, and Peter Rhodes. Investigation of large display color image appearance-III: Modeling image naturalness. *JIST*, 53(3):31104–1, 2009.
- [46] Anuj Srivastava, Ann B Lee, Eero P Simoncelli, and Song-Chun Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18(1):17–33, 2003.
- [47] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.
- [48] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.
- [49] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2178. IEEE, 2006.
- [50] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.
- [51] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
- [52] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005.
- [53] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.
- [54] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.
- [55] Adam Coates, Andrew Y Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International conference on artificial intelligence and statistics*, pages 215–223, 2011.
- [56] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [57] Rachel McDonnell, Martin Breidt, and Heinrich H Bühlhoff. Render me real?: investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG)*, 31(4):91, 2012.
- [58] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.