Some like it hot - visual guidance for preference prediction

Rasmus Rothe
CVL, ETH Zurich
rrothe@vision.ee.ethz.ch

Radu Timofte
CVL, ETH Zurich
radu.timofte@vision.ee.ethz.ch

Luc Van Gool
KU Leuven, ETH Zurich
vangool@vision.ee.ethz.ch

Abstract

For people first impressions of someone are of determining importance. They are hard to alter through further information. This begs the question if a computer can reach the same judgement. Earlier research has already pointed out that age, gender, and average attractiveness can be estimated with reasonable precision. We improve the state-of-the-art, but also predict - based on someone's known preferences - how much that particular person is attracted to a novel face. Our computational pipeline comprises a face detector, convolutional neural networks for the extraction of deep features, standard support vector regression for gender, age and facial beauty, and - as the main novelties - visual regularized collaborative filtering to infer interperson preferences as well as a novel regression technique for handling visual queries without rating history. We validate the method using a very large dataset from a dating site as well as images from celebrities. Our experiments yield convincing results, i.e. we predict 76% of the ratings correctly solely based on an image, and reveal some sociologically relevant conclusions. We also validate our collaborative filtering solution on the standard MovieLens rating dataset, augmented with movie posters, to predict an individuals movie rating. We demonstrate our algorithms on howhot.io *which went viral around the Internet with more than 50 million pictures evaluated in the first month.*

1. Introduction

'First impressions count' the saying goes. Indeed, psychology confirms that it only takes 0.1 seconds for us to get a first impression of someone [47], with the face being the dominant cue. Factors that are relevant for survival seem the ones evolution makes us pick up the fastest. These include age, gender, and attractiveness. We will call those quantifiable properties, such as age and gender, 'demographics'.
An everyday demonstration is that people on dating sites often base their decisions mainly on profile images rather than textual descriptions of interests or occupation. Our goal is to let a computer predict someone's preferences, from single facial photos (in the wild). In particular, we try to predict how much a previously unseen face would be attractive for
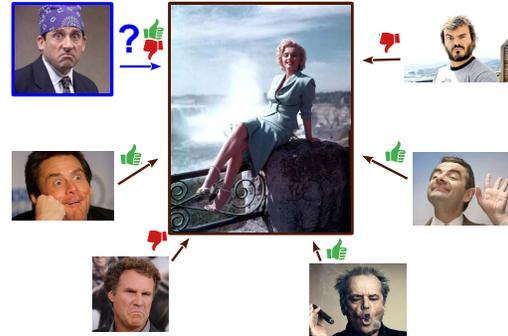

Figure 1. Can we infer preferences from a single image?

a particular person who has already indicated preferences for people in the system.

Our main benchmark is a large dataset of more than 13,000 user profiles from a dating site. We have access to their age and gender, as well as more than 17 million 'hot or not' ratings by some users about some other users (their profile photo). The ratings are very sparse when compared to their potential number. For people who have given ratings, we want to predict new ratings for other people in and outside the dataset.

The visual information, here the profile image, presumably containing a face, is the main basis for any user-to-user rating. Therefore, we employ a face detector and crop the best detected face and its surrounding context (corresponding to body and posture) from which we extract deep features by means of a (fine-tuned) convolutional neural network. In order to make sure that these features are appropriate for the main task - automated attractiveness rating - we first test our features on age, gender, and facial beauty estimation for which previous methods and standard datasets exist. We obtain state-of-the-art results.

For predicting preferences for users with known ratings for a subset of others in the dataset, collaborative filtering is known to provide top results, *i.e.* for movie [22] or advertisement suggestions [36]. We adapt this framework to take account of visual information, however. As our experiments will show, adding visual information improves the prediction, especially in cases with few ratings per user. In case of a new face, not part of the dataset and thus without a history of preferences, we propose to regress the input im-

age to the latent space of the known users. By doing so, we alleviate the need for past ratings for the query and solely rely on the query image.

The same technique can be applied to different visuals-enhanced tasks, such as rating prediction of movies, songs, shopping items, in combination with a relevant image (*e.g.* movie poster, CD cover, image of the item). We test on the MovieLens dataset augmented with poster images for each movie, a rather weak information, to demonstrate the wider applicability of our approach.

We demonstrate our algorithms on `howhot.io`, a website where people can upload a photo of their face and an algorithm will then estimate the age, gender and facial attractiveness of the person.

The main contributions of our work are:

- an extensive study on the inference of information from profile images using the largest dating dataset thus far
- a novel collaborative filtering approach that includes visual information for rating/preference prediction
- a novel regression technique for handling visual queries without rating history which prior work cannot cope with

2. Related Work

The focus of our paper is to infer as much information as possible from a single image and to predict subjective preferences based on an image query with possibly a prior rating history. Next we review related works.

Image features. Instead of handcrafted features like SIFT, HoG, or Gabor filters, we use learned features obtained using neural networks [12, 23, 40]. The latter have shown impressive performance in recent years. Such features have already been used for age and gender estimation in [37, 45].

Demographics estimation. Multiple demographic properties such as age, gender, and ethnicity have been extracted from faces. A survey on age prediction is provided by Fu *et al.* [10] and on gender classification by Ng *et al.* [33]. Kumar *et al.* [25] investigate image 'attribute' classifiers in the context of face verification. Some approaches need face shape models or facial landmarks [18, 20], others are meant to work in the wild [5, 6, 37, 45] but still assume face localization. Generally, the former approaches reach better performance as they use additional information. The errors in model fitting or landmark localization are critical. Moreover, they require supervised training, detailed annotated datasets, and higher computation times. On top of the extracted image features a machine learning approach such as SVM [43] is employed to learn a demographics prediction model which is then applied to new queries.

Subjective property estimation. While age and gender correspond to objective criteria, predicting the attractiveness of a face is more subjective. Nonetheless, facial beauty [1, 8, 13, 32, 48] can still be quantified by averaging the ratings by a large group. Benchmarks and corresponding estimation methods have been proposed. In the subjective direction, Dhar *et al.* [7] demonstrate the aesthetic quality estimation and predict what they call 'interestingness' of an image, while Marchesotti *et al.* [30] discover visual attributes (including subjective ones) to then to use them for prediction. Also, recently Kiapour *et al.* [21] inferred complex fashion styles from images. Generally, the features and methods used for age and gender can be adapted to subjective property estimation, and we do the same in this paper. From the literature we can observe three trends: (i) besides Whitehill and Movellan [46], most papers focus on predicting facial beauty averaged across all ratings, whereas we aim at predicting the rating by a specific person; (ii) as pointed out in the study by Laurentini and Bottino [26] usually small datasets are used, sometimes with less than 100 images and with only very few ratings per image – our dataset contains more than 13,000 images with more than 17 million ratings; (iii) most datasets are taken in a constrained environment showing aligned faces. In contrast, our photos contain in many cases also parts of the body and some context in the background. Thus, we focus not just on facial beauty but on general attractiveness of the person – referred to as hotness in the following.

Preferences/ratings prediction. The Internet brought an explosion of choices. Often, it is difficult to pick suitable songs to hear, books to read, movies to watch, or - in the context of dating sites - persons to contact. Among the best predictors of interest are collaborative filtering approaches that use the knowledge of the crowd, *i.e.* the known preferences/ratings of other subjects [2, 38]. The more prior ratings there are, the more accurate the predictions become. Shi *et al.* [39] survey the collaborative filtering literature. Matrix factorization lies at the basis of many top collaborative filtering methods [22, 27]. Given the importance of the visual information in many applications, we derive a matrix factorization formulation regularized by the image information. While others [39] proposed various regularizations, we are the first to prove that visual guidance helps preference prediction. Moreover, we propose to regress queries without rating history to a latent space derived through matrix factorization for the known subjects and ratings.

Social networks. The expansion of the internet and the advance of smartphones boosted the (online) social networks worldwide. Networks such as Facebook facilitate interaction, sharing, and display of information and preferences among individuals. Yet, time is precious and hence efforts are made to develop filters and ranking tools to assist users. A recent study by Youyou *et al.* [49] shows that accurate predictions about the personality of a user can be made using her/his 'likes'. Contents and ads can then be personalized and this is extremely important for social networks
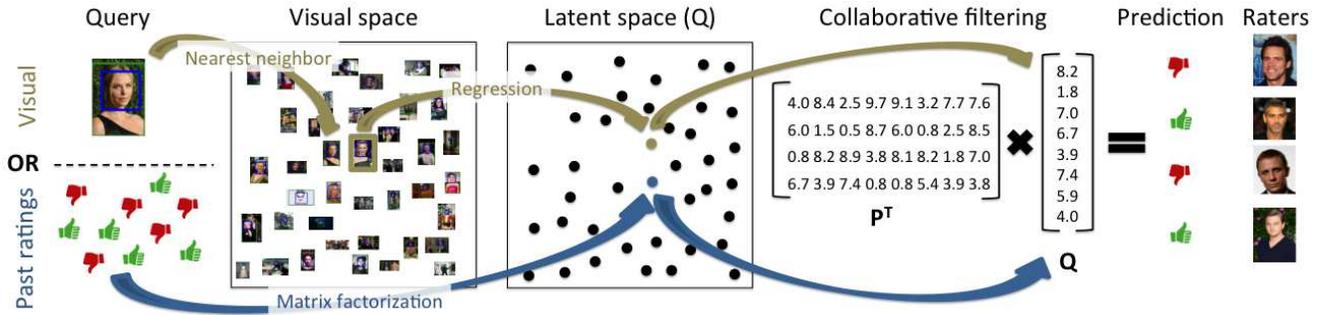
Figure 2. Preferences prediction scheme. For a visual query without past ratings we first regress to the latent Q space (obtained through matrix factorization) to then obtain the collaborative filtering prediction as in the case for the queries with known past ratings and Q factor.

and search engines such as Google [41]. This paper focuses on dating sites and the prediction of attractiveness ratings. Most such works [3, 24] rely on past ratings and cannot cope when there are none or few.

3. Visual features

Razavian *et al*. [34] showed that features extracted from convolutional neural networks (CNN) are very powerful generic descriptors. Inspired by that, for all our experiments we use the VGG-16 [40] features which are pre-trained on a large ImageNet object dataset and result in a descriptor of length 4,096. We use the implementation by Vedaldi and Lenc [44]. We reduce the dimensionality using PCA to keep $\sim 99\%$ of the energy. Before we use these feature to predict attractiveness to a particular user, we first confirm that the extracted visual features are powerful enough to capture minor facial differences by predicting age and gender.

We perform reference experiments on a widely used dataset for age prediction, the MORPH 2 database [35]. We also test gender estimation on the same MORPH 2 dataset. Unfortunately, besides the dataset provided by Gray *et al*. [13] – to the best of our knowledge – there are no other publicly available large datasets on averaged facial beauty. As shown next our features achieve state-of-the-art performance for age, gender, and facial beauty prediction. We believe that this good performance is mostly due to the depth of the model with 16 layers, compared with previous state-of-the-art using only 6 layers [45].

3.1. Predicting age and gender

Our experiments are conducted on a publicly available dataset, the MORPH 2 database [35]. We adopt the experimental setup of [5, 6, 15, 45], where a set of 5,475 individuals is used whose age ranges from 16 to 77. The dataset is randomly divided into 80% for training and 20% for testing. Following the procedure described in [12], our CNN features are fine-tuned on the training set.

The age is regressed using Support Vector Regression (SVR) [4] with an RBF kernel and its parameters are set by cross-validation on a subset of the training data. We report the performance in terms of mean absolute error (MAE) be-

tween the estimated and the ground truth age.

Method	MORPH 2 [35]
AGES [11]	8.83
MTWGP [50]	6.28
CA-SVR [6]	5.88
SVR [15]	5.77
OHRank [5]	5.69
DLA [45]	4.77
Proposed Method	**3.45**

Table 1. Age estimation performance in terms of mean absolute error (MAE) on the MORPH 2 dataset. We improve the state-of-the-art results by more than 1 year.

Figure 3. Average faces for 5 clusters based on age or beauty, resp. Average beauty is less meaningful, suggesting personalized prediction.

As shown in Table 1, we achieve state-of-the-art performance on the MORPH 2 dataset (3.45 years MAE) by reducing the error of the currently best result (4.77 MAE reported by [45]) with more than a full year. For gender prediction on the MORPH 2 dataset we keep the same partition as for age and achieve 96.26% accuracy, which, despite the small training set, is on par with other results in the literature [16, 17]. Fig. 4 shows several good and erroneous predictions of our method on the MORPH 2 dataset. Fig. 3 shows averages of faces ranked according to age on MORPH 2 and beauty on Gray, resp. On our more challenging Hot-or-Not dataset (Section 5.1.1) we achieve 3.61 MAE for age and 88.96% accuracy for gender prediction.

3.2. Predicting facial beauty

Following a similar procedure as for age prediction, we test our features on the dataset introduced by Gray *et al*. [13]. The Gray dataset contains 2056 images with female faces collected from a popular social/dating website[1]. The facial beauty was rated by 30 subjects and the ratings were then normalized as described in [13]. The dataset is split into 1028 images for training and 1028 for testing. We

---

[1]http://hotornot.com

report the average performance across exactly the same 5 splits from the reference paper in terms of Pearson's correlation coefficient, the metric from the original paper. Also, we report performance with and without face alignment using the same alignment algorithm of Huang *et al.* [19].

Method	Correlation w/o alignment	Correlation w/ alignment
Eigenface	0.134	0.180
Single Layer Model [13]	0.403	0.417
Two Layer Model [13]	0.405	0.438
Multiscale Model [13]	0.425	0.458
Proposed Method	**0.470**	**0.478**

Table 2. Facial beauty estimation performance on Gray dataset with and without face alignment in terms of correlation.

As shown in Table 2 our proposed features achieve state-of-the-art performance on predicting facial beauty as averaged over multiple raters. We improve by more than 10% over the best score reported by [13] for the raw images. A couple of per image results are depicted in Fig. 4.

4. Predicting preferences

Our goal is to make personalized predictions, such as how a specific male subject $m \in M$ rates a female subject $f \in F$. The rating R_{mf} is 1 if 'm likes f', -1 if 'm dislikes f', and 0 if unknown. f is also called the query user, as at test time we want to predict the individual ratings of all men for that woman. Due to space limitations, we derive the formulation for this case. Yet it is also valid when swapping sexes, *i.e.* when women are rating men.

In the following section we phrase the problem as a collaborative filtering problem, assuming that we know past ratings for both men and women. In Section 4.2 we extend the formulation to also consider the visuals of the subjects being rated. In Section 4.3 we present a solution to predict the ratings solely based on the visual information of the subjects, without knowing how they were rated in the past.

4.1. Model-based collaborative filtering (MF)

We phrase the problem of a user m rating the image of user f as a model-based collaborative filtering problem. The model learned from known ratings is then used to predict unknown ratings. In its most general form, we have

$$g(P_m, Q_f) \Rightarrow R_{mf}, \ m = 1, 2, ..., M, \ f = 1, 2, ..., F, \quad (1)$$

where the function g maps the model parameters to the known ratings. P_m denotes a set of model parameters describing the preferences of user m. Similarly, Q_f describes the appearance of user f, *i.e.* a low-dimensional representation of how the appearance of a user is perceived. We now estimate the model parameters given the ratings we know.

In recent years, Matrix Factorization (MF) techniques have gained popularity, especially through the Netflix challenge, where it achieved state-of-the-art performance [22]. The basic assumption underlying MF models is that we can learn low-rank representations, so-called latent factors, to predict missing ratings between user m and image f. One can approximate the ratings as

$$R \approx P^T Q = \hat{R}. \quad (2)$$

In the most common formulation of MF [39] we can then frame the minimization as

$$P^*, Q^* = \underset{P,Q}{\operatorname{argmin}} \frac{1}{2} \sum_{m=1}^{M} \sum_{f=1}^{F} I_{mf}(R_{mf} - P_m^T Q_f)^2 + \frac{\alpha}{2}(\|P\|^2 + \|Q\|^2) \quad (3)$$

where P and Q are the latent factors and P^* and Q^* their optimal values. I_{mf} is an indicator function that equals 1 if there exists a rating R_{mf}. The last term regularizes the problem to avoid overfitting.

4.2. Visual regularization (MF+VisReg)

Knowing that the users in the app rate the subjects of the opposite sex solely based on the image, we make the assumption that people with similar visual features have similar latent appearance factors Q. Thus we can extend the formulation by adding the visual features V of the query images to further regularize the optimization

$$L(P, Q) = \frac{1}{2} \sum_{m=1}^{M} \sum_{f=1}^{F} I_{mf}(R_{mf} - P_m^T Q_f)^2 + \frac{\alpha_1}{2}(\|P\|^2 + \|Q\|^2) + \frac{\alpha_2}{2} \sum_{f=1}^{F} \sum_{g=1}^{F} (S_{fg} - Q_f^T Q_g)^2. \quad (4)$$

The visual similarity is defined as

$$S_{fg} = \frac{V_f^T V_g}{\|V_f\| \|V_g\|}. \quad (5)$$

Visually this proves to be a good metric for visual similarity. The optimal latent factors are calculated by gradient descent, where the derivatives are

$$\frac{\partial L}{\partial P_m} = \sum_{f=1}^{F} I_{mf}(P_m^T Q_f - R_{mf})Q_f + \lambda P_m$$
$$\frac{\partial L}{\partial Q_f} = \sum_{m=1}^{M} I_{mf}(P_m^T Q_f - R_{mf})P_m \quad (6)$$
$$+ 2\alpha_2 \sum_{g=1}^{F} (Q_f^T Q_g - S_{fg})Q_g + \lambda Q_f.$$

4.3. Visual query

We now want to predict how user m rates user f without knowing any past ratings of f but knowing her visual feature V_f (see Fig. 2). This implies that we do not know the latent factor Q_f for f. The goal is to get an estimate $\hat{Q}_f$ of Q_f based solely on the visual feature V_f. Then we would be able to regress the rating as

$$\hat{R}_{mf} = P_m^T \hat{Q}_f. \quad (7)$$

Learning a global regression led to poor results as attractiveness is highly subjective. Instead our approach is inspired by the recently introduced anchored neighborhood regression (ANR) method for image super-resolution [42], where

	MORPH 2 (age)						MORPH 2 (gender)				Gray (facial beauty)					
Prediction	57.0	41.0	29.0	31.1	35.4	39.4	Male	Female	Male	Female	-0.88	-0.58	0.80	0.21	-0.56	0.14
Ground truth	57	41	29	45	55	23	Male	Female	Female	Male	-0.96	-0.59	0.79	2.68	2.60	-2.89

Figure 4. Examples of accurately and wrongly predicted age, gender, and facial beauty for the MORPH 2 and Gray datasets.

the problem is formulated as a piece-wise local linear regression of low to high resolution image patches and with offline trained regressors. In contrast to ANR, each sample is an anchor and the neighborhood is spanned over all other training samples and weighted according to its similarity to the anchor. This way we are obtaining more robust local regressors that can cope with the scarcity of the data.

As for regularizing MF, we assume that the visual features V and the latent factor Q locally have a similar geometry. Further, we assume that we can locally linearly reconstruct each visual feature or latent factor by its neighbors. Under these assumptions we can reconstruct features and latent factors using the same weights for the neighbors. In the visual space we now aim to find these weights β by phrasing the problem as a ridge regression

$$\min_{\beta_g}\|V_g - N_{V_g}\beta_g\|^2 + \lambda\left(\kappa\|\Gamma_g\beta_g\|^2 + (1-\kappa)\|\beta_g\|^2\right), \quad (8)$$

where N_{V_g} is a matrix of the neighboring visual features of V_g stacked column-wise and κ is a scalar parameter. The optimization is regularized by the similarity to its neighbors according to eq. 5, in the sense that greater similarity yields greater influence on β:

$$\Gamma_g = \mathrm{diag}(1 - S_{g1}, 1 - S_{g2}, ..., 1 - S_{gF}). \quad (9)$$

The closed-form solution of the problem can be written as

$$\beta_k = \left[N_{V_g}^T N_{V_g} + \lambda\left[\kappa\Gamma_g^T\Gamma_g + (1-\kappa)I\right]\right]^{-1} N_{V_g}^T V_g. \quad (10)$$

As we assume that the latent space behaves similarly locally, we can regress the latent factor Q_g as a linear combination of its neighbors using the same β_g. Note that N_{Q_g} corresponds to the latent factors of N_{V_g}, i.e. the neighbors in the visual space. Plugging in our solution for β_g we get

$$\begin{aligned} Q_g &= N_{Q_g}\beta_g \\ &= N_{Q_g}\left[N_{V_g}^T N_{V_g} + \lambda\left[\kappa\Gamma_g^T\Gamma_g + (1-\kappa)I\right]\right]^{-1} N_{V_g}^T V_g \quad (11) \\ &= M_g V_g. \end{aligned}$$

Thus we have found a projection M_g from a visual feature V_g to its latent factor Q_g. At test time for a given visual feature V_f, we now aim to find the most similar visual feature in the training space, $\hat{g} = \arg\max_g S_{fg}$. Then we use the projection matrix of $\hat{g}$ to obtain $\hat{Q}_f$ to finally estimate the rating of user m for the image of user f as

$$\hat{R}_{mf} = P_m^T\hat{Q}_f, \quad \hat{Q}_f = M_{\hat{g}}V_f. \quad (12)$$

5. Experiments

In this section we present qualitative and quantitative results of our proposed framework on the Hot-or-Not and the MovieLens dataset.

5.1. Hot-or-Not

5.1.1 The dataset

Our dataset was kindly provided by Blinq[2], a popular hot-or-not dating application. We will make the anonymized ratings and visual features of the last layer available under http://www.vision.ee.ethz.ch/~rrothe/. The app shows the user people of the sex of interest, one after the other. The user can then like or dislike them. If both like each other's profile photo they are *matched* and can chat to each other. People can select up to 5 photos from Facebook for their profile.

Dataset statistics. Before performing any experiments we removed underage people, anyone over 37 and bi- and homosexual users as these comprise only a small minority of the dataset. All users who received less than 10 ratings were also removed. As the majority of the people decide on the first photo, we ignore the other photos. The resulting dataset has 4,650 female users and 8,560 male users. The median age is 25. In total there are 17.33 million ratings, 11.27m by men and 6.05m by women. Interestingly, 44.81% of the male ratings are positive, while only 8.58% of the female ratings are positive. Due to this strong bias of ratings by women, we only predict the ratings of men. There are 332,730 matches.
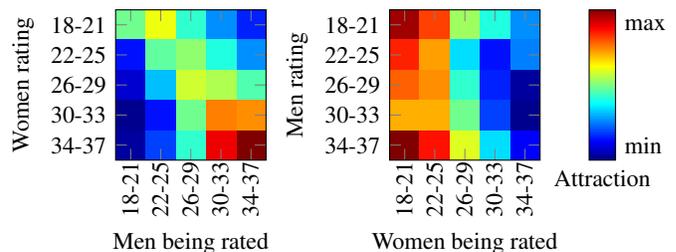

Figure 5. Preferences by age for women and men.

Preferences bias. To investigate the natural bias caused by age, we divide the men and women from our dataset according to their age and gender. For each age group of men we counted the percent of hot vs. not on the ratings towards the age groups of women and vice versa. Fig. 5 describes the preferences by age as found in our dataset.

---

[2]www.blinq.ch

Women generally prefer slightly older men and give better ratings the older they get. In comparison, men on this app prefer women under 25.
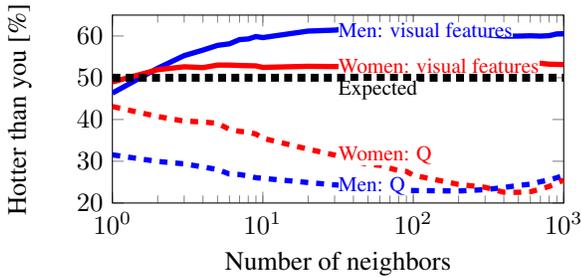

Figure 6. **Hotness paradox.** The people visually similar to you are on average hotter than you. The situation changes when we compute the similarity based on learned latent Q representations.

Hotness paradox. We notice an interesting phenomenon. Most people have a lower rating than their visually similar neighbors, on average, and this holds for both men and women. In Fig. 6 we report the percentage of cases where the average hotness of the neighborhood of a subject is greater than that of the subject, and this for a neighborhood size from 1 to 10^3. We plot also the results when we use our latent Q representations for retrieving neighbors. Surprisingly, this time the situation is reversed, the subjects tend to be hotter than their Q-similar neighborhood. Regardless the choice of similarity we have a strong deviation from the expected value of 50%. We call this phenomenon the 'Hotness paradox'. It relates to the so-called 'Friendship paradox' [9] in social networks, where most people have fewer friends than their friends have.

Visual features. As the space covered by the person varies greatly between images, we run a top face detector [31] on each image. Then we crop the image to the best scoring face detection and include its surrounding (100% of the width to each side and 50% of the height above and 300% below), to capture the upper-body and some of the background. If the face is too small or the detection score too low, we take the entire image. Then we extract CNN features.

5.1.2 Experimental setup

For all experiments, 50% of either gender are used for training and the rest for testing. For each user in the testing set, 50% of the received ratings are used for testing. We compare different methods. *Baseline* predicts the majority rating in the training set. Matrix factorization is applied without and with visual regularization, *MF* ($\alpha_1 = 0.1$) and *MF+VisReg* ($\alpha_2 = 0.1$), resp. The dimensionality of the latent vector of P and Q is fixed to 20. The other parameters were set through cross-validation on a subset of the training data. We predict the ratings a subject receives based upon different knowledge: For *Visual* we solely rely on the image of the subject which means that we do not know any ratings the subject has received so far. For *10 Ratings, 100 Ratings,*

Full History, we instead base the prediction upon a fixed set of known ratings for each query user. We report the average accuracy, *i.e.* the percentage of correctly predicted ratings of the testing set, and the Pearson's correlation.

5.1.3 Results

Performance. Fig. 7 shows how the average accuracy varies with the number of known past ratings for the query user. We report the average performance across all men's ratings. Knowing just the image of the person, we can predict 75.92% of the ratings correctly. Adding past received ratings of the user improves performance to up to 83.64%. Matrix factorization significantly improves as more ratings are known. If only few ratings are available, regularizing the matrix factorization with the visuals boosts performance significantly, *i.e.* from 72.92% to 78.68% for 10 known ratings. Table 3 summarizes the results for various settings.
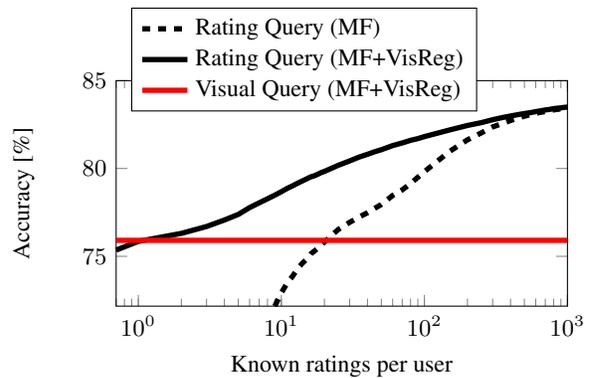

Figure 7. Number of known ratings for a female query user vs. accuracy of predicted male's ratings.

	Query	Accuracy	Correlation
Baseline	N/A	54.92%	N/A
MF	Visual	75.90%	0.520
MF+VisReg		**75.92%**	**0.522**
MF	10 Ratings	72.92%	0.456
MF+VisReg		78.68%	0.576
MF	100 Ratings	79.82%	0.593
MF+VisReg		81.82%	0.635
MF	Full History	83.62%	0.671
MF+VisReg		**83.64%**	**0.671**

Table 3. Preference prediction results on Hot-or-Not dataset for female queries.

Latent space Q vs. preferences. In Fig. 9 we show the learned latent space Q from the matrix factorization by PCA projecting it to two dimensions and adding the hotness and age properties for both genders with visual regularization. The learned latent factor Q captures appearance and for women there is a clear separation in terms of attractiveness and age, whereas for men the separation is less obvious.

According to the preferences P and the learned latent Q one can have a more in-depth view on the appearance of women and men. In Fig. 10 both men and women are clustered according to their 2D representation of the learned latent factors P (preferences of men) and Q (appearances of

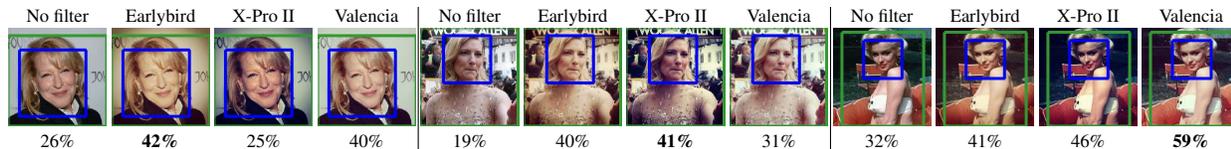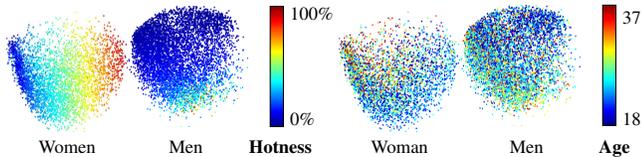| No filter	Earlybird	X-Pro II	Valencia	No filter	Earlybird	X-Pro II	Valencia	No filter	Earlybird	X-Pro II	Valencia
26%	**42%**	25%	40%	19%	40%	**41%**	31%	32%	41%	46%	**59%**

Figure 8. Improving the hotness rating by Instagram filters.

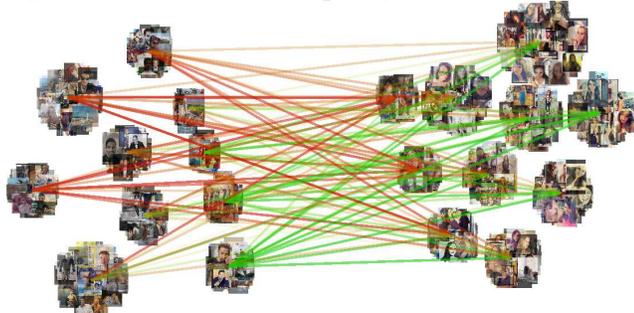Figure 9. Visualization of latent space Q for women and men.

Figure 10. Preferences between clusters of users. The color of the arrow indicates how much the men's cluster likes (green) or dislikes (red) the women's cluster on average.

women), respectively. For visualization purposes we used 100 user images for each cluster and 10 clusters. The men are visually more diverse in each of their clusters than the women in their clusters, because the men are clustered according to their preferences, therefore ignoring their visual appearance, while the women are clustered according to their Q factors which are strongly correlated with appearance and hotness, as shown in Fig. 9.

Visual queries without past ratings. We validate our approach on images outside our dataset, retrieved from the internet for celebrities. By applying the visual query regression to the Q space we can make good predictions for such images. For a visual assessment see Fig. 12. This figure also depicts a number of issues our pipeline faces: too small faces, detector failure, wrongly picked face, or simply a wrong prediction. We also tested our method on cartoons and companion pets with the predictor trained on Hot-or-Not. The results are surprising.

Instagram filters or how to sell your image. Images and their hotness prediction also indicate which changes could improve their ratings. Earlier work has aimed at the beautification of a face image by invasive techniques such as physiognomy changes [28] or makeup [14, 29]. Yet, we found that non-invasive techniques (not altering facial geometry and thus 'fair') can lead to surprising improvements. We have evaluated the most popular Instagram filters[3] for our

---
[3]brandongaille.com/10-most-popular-instagram-photo-filters

task. We observed that the filters lead to an increase in predicted hotness. In Fig. 8 we show a couple of results in comparison to the original image. Note that with our predictor and such Instagram filters a user can easily pick its best profile photo.

5.1.4 howhot.io

We demonstrate our algorithms on howhot.io, a website where people can upload a photo of their face and our algorithm will then estimate the age, gender and facial attractiveness of the person (c.f. Fig. 11). The CNN was trained on the Hot-or-Not dataset for predicting attractiveness and on the IMDB-WIKI dataset [37]

Figure 11. howhot.io

for age and gender prediction. The website went viral around the Internet with more than 50 million pictures evaluated in the first month.
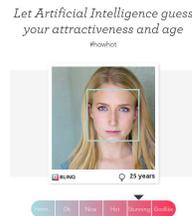
5.2. MovieLens

5.2.1 The dataset

We also perform experiments on the MovieLens 10M[4] dataset. It contains 10,000,054 ratings from 69,878 users for 10,681 movies. Ratings are made on a 5-star scale, with half-star increments. On average, each movie has 3,384 ratings and each user rates 517 movies. Note that even though there are more than 10 million ratings, the rating matrix is sparse with only 1.34% of all ratings known. We augment each movie with the poster image from IMDB and extract the same deep CNN features as for the Hot-or-Not dataset. We will make the poster images publicly available under http://www.vision.ee.ethz.ch/~rrothe/.

5.2.2 Experimental Setup

The experimental setup in term of training and testing split is identical to the Hot-or-Not dataset. As the movie posters are much less informative regarding the ratings in comparison to the Hot-or-Not images, the visual regularization is reduced to $\alpha_2 = 0.001$. For a given movie we want to infer the ratings of all users. Again, we evaluate the case where just the poster is known and also cases where a varying number of ratings is known. As a baseline we show how a prediction made at random would perform, assuming that there is no bias in the ratings of the test set.
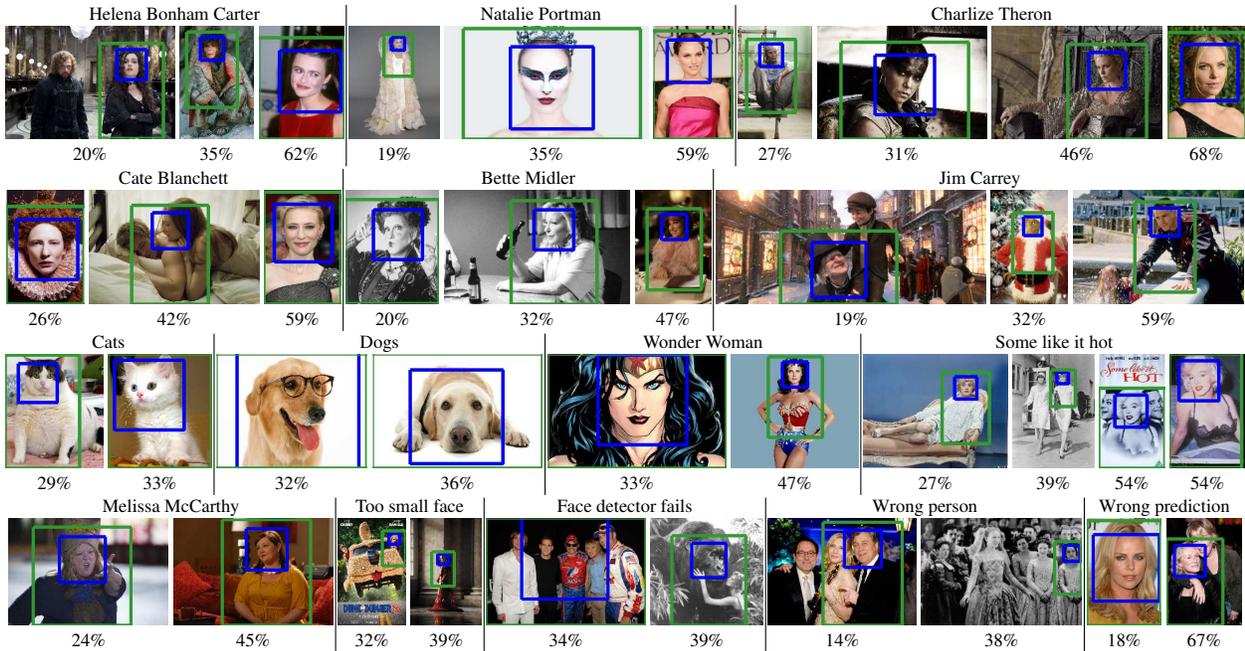
---
[4]grouplens.org/datasets/movielens

Figure 12. Predicted percentage of positive ratings for numerous celebrities by the user base of the Hot-or-Not dataset.

	Correctly predicted						Overrated poster			Underrated poster		
Average prediction	3.9	4.0	3.1	3.5	2.7	2.4	3.8	3.5	3.7	2.4	2.9	2.7
Average rating	4.1	4.0	3.1	3.5	2.7	2.5	0.6	1.4	1.9	3.9	3.9	4.5

Figure 13. Examples of predicted ratings for various movie posters solely based on the visual information of the poster.

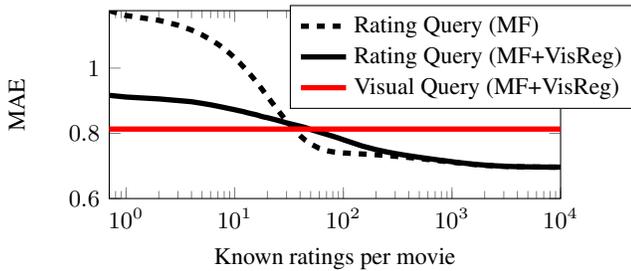Figure 14. Number of known ratings for a movie vs. MAE of the predicted ratings.

	Query	MAE	Correlation
Baseline	N/A	1.507	N/A
MF	Visual	0.824	0.286
MF+VisReg		**0.813**	**0.292**
MF	10 Ratings	1.031	0.280
MF+VisReg		0.872	0.270
MF	100 Ratings	0.740	0.467
MF+VisReg		0.780	0.461
MF	Full History	0.696	0.530
MF+VisReg		**0.696**	**0.536**

Table 4. Rating prediction results on augmented MovieLens.

5.2.3 Results

Table 4 summarizes the performance. Fig. 14 shows how the number of known ratings impacts the MAE. Visual regularization of MF improves performance, especially when few ratings are known, *i.e.* for 10 known ratings the MAE can be reduced by 15% from 1.031 to 0.872. When just the movie poster is known, the MAE is 0.813, which is on par with knowing 30 ratings. Fig. 13 shows some movie posters. We also show overrated and underrated posters, *i.e.* posters where our algorithm - based on the poster - predicts a much better or worse score than the actual movie rating.

6. Conclusion

We proposed a collaborative filtering method for rating/preference prediction based not only on the rating history but also on the visual information. Moreover, we can accurately handle queries with short or lacking rating history. We evaluated our system on a very large dating dataset and on the MovieLens dataset augmented with poster images. To the best of our knowledge, we are the first to report on such a large dating dataset, and to show that adding weak visual information improves the rating prediction of collaborative filtering methods on MovieLens. We achieved state-of-the-art results on facial beauty, age and gender prediction and give some sociologically interesting insights.

References

[1] H. Altwaijry and S. Belongie. Relative ranking of facial attractiveness. In *WACV*, 2013.

[2] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *UAI*, 1998.

[3] L. Brozovsky and V. Petricek. Recommender system for online dating service. *arXiv preprint cs/0703042*, 2007.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *TIST*, 2:1–27, 2011.

[5] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR*, 2011.

[6] K. Chen, S. Gong, T. Xiang, and C. C. Loy. Cumulative attribute space for age and crowd density estimation. In *CVPR*, 2013.

[7] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, 2011.

[8] Y. Eisenthal, G. Dror, and E. Ruppin. Facial attractiveness: Beauty and the machine. *Neural Computation*, 18(1):119–142, 2006.

[9] S. L. Feld. Why your friends have more friends than you do. *American Journal of Sociology*, 96(6):1464–1477.

[10] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. *PAMI*, 32(11):1955–1976, 2010.

[11] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *PAMI*, 29(12):2234–2240, 2007.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[13] D. Gray, K. Yu, W. Xu, and Y. Gong. Predicting facial beauty without landmarks. In *ECCV*. 2010.

[14] D. Guo and T. Sim. Digital face makeup by example. In *CVPR*, 2009.

[15] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *TIP*, 17(7):1178–1188, 2008.

[16] G. Guo and G. Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *CVPR*, 2011.

[17] G. Guo and G. Mu. Joint estimation of age, gender and ethnicity: Cca vs. pls. In *FG*, 2013.

[18] H. Han, C. Otto, and A. Jain. Age estimation from face images: Human vs. machine performance. In *ICB*, 2013.

[19] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007.

[20] A. Jaimes and N. Sebe. Multimodal human–computer interaction: A survey. *CVIU*, 108(1):116–134, 2007.

[21] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*. 2014.

[22] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[24] A. Krzywicki, W. Wobcke, X. Cai, A. Mahidadia, M. Bain, P. Compton, and Y. S. Kim. Interaction-based collaborative filtering methods for recommendation in online dating. In *WISE*. 2010.

[25] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[26] A. Laurentini and A. Bottino. Computer analysis of face beauty: A survey. *CVIU*, 125:184–199, 2014.

[27] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2001.

[28] T. Leyvand, D. Cohen-Or, G. Dror, and D. Lischinski. Data-driven enhancement of facial attractiveness. *SIGGRAPH*, 2008.

[29] L. Liu, J. Xing, S. Liu, H. Xu, X. Zhou, and S. Yan. Wow! you are so beautiful today! *TOMM*, 11:1–22, 2014.

[30] L. Marchesotti, N. Murray, and F. Perronnin. Discovering beautiful attributes for aesthetic image analysis. *IJCV*, 113(3):246–266, 2015.

[31] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*. 2014.

[32] Y. Mu. Computational facial attractiveness prediction by aesthetics-aware features. *Neurocomputing*, 99:59–64, 2013.

[33] C. B. Ng, Y. H. Tay, and B.-M. Goi. Recognizing human gender in computer vision: a survey. In *PRICAI*. 2012.

[34] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR Workshops*, 2014.

[35] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *FG*, 2006.

[36] G. B. Robinson. Automated collaborative filtering in world wide web advertising, 1999. US Patent 5,918,014.

[37] R. Rothe, R. Timofte, and L. Van Gool. Dex: Deep expectation of apparent age from a single image. In *ICCV Workshops*, 2015.

[38] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW*, 2001.

[39] Y. Shi, M. Larson, and A. Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *CSUR*, 47(1):3, 2014.

[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *WWW*, 2004.

[42] R. Timofte, V. De Smet, and L. Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *ACCV*, 2014.

[43] V. N. Vapnik. *Statistical Learning Theory*. 1998.

[44] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. *arXiv preprint arXiv:1412.4564*, 2014.

[45] X. Wang, R. Guo, and C. Kambhamettu. Deeply-learned feature for age estimation. In *WACV*, 2015.

[46] J. Whitehill and J. R. Movellan. Personalized facial attractiveness prediction. In *FG*, 2008.

[47] J. Willis and A. Todorov. First impressions: Making up your mind after 100 ms exposure to a face. *Psychological Science*, 17:592–598, 2006.

[48] H. Yan. Cost-sensitive ordinal regression for fully automatic facial beauty assessment. *Neurocomputing*, 129:334–342, 2014.

[49] W. Youyou, M. Kosinski, and D. Stillwell. Computer-based personality judgments are more accurate than those made by humans. *National Academy of Sciences*, 2015.

[50] Y. Zhang and D.-Y. Yeung. Multi-task warped gaussian process for personalized age estimation. In *CVPR*, 2010.