Extended Features using Machine Learning Techniques for Photo Liking Prediction

Steve Göring, Konstantin Brand, Alexander Raake
Dept. of Audio Visual Technology; Technische Universitt Ilmenau, Germany
Email: [steve.goering, konstantin.brand, alexander.raake]@tu-ilmenau.de

***Abstract*—Today several photo platforms provide thousands of new pictures, it becomes ambitious to find highly appealing or like-able photos within such loads of data. Here, automatic liking prediction can support users in handling their pictures or improve ranking in sharing platforms. We describe a machine learning approach for photo liking prediction. Our features are based on various techniques, e.g. natural language processing/sentiment analysis, pre-trained deep learning networks, social network analysis and extended previously reported features. We conduct large-scale experiments using a collected dataset consisting of 80k photos based on two main categories from 500px with different settings. In our experiments we analyzed the impact of our newly features and found that social network features have the strongest influence for liking prediction, we achived a boost of 15%. Furthermore, we show that all implemented features are able to improve prediction accuracy of liking rates. We additionally analyze which groups of features that can be derived directly from pictures are usable for prediction.**

I. Introduction

Today thousands of pictures are uploaded every hour on photo-sharing platforms[1], such as Flickr, 500px, DeviantArt, Instagram, because it is easy to take and share pictures there. You can use your smart-phone's camera in various situations or places, and with the ease of the mobile internet, uploading and sharing can be done within seconds. Considering the enormous amount of new photos it is more and more important to filter and rate uploaded photos based on their appeal or expected liking. For a user it is crucial to know which of the uploaded photos are of high appeal and will most probably be liked by other people. In general, the image liking in the context of a photo sharing site is based on different types of factors, for example perspective, lightning, colors, subject and framing [4]. Also, other factors influence liking, e.g., user preferences, photo category, position and reputation of the user and the spectators. Photo-sharing platforms also include typical view- and liking-counters, category tagging, and a commenting section. Mostly they are combined with a social network, in order to let users form similar interest groups where they can exchange their knowledge with each other.

After uploading a photo to such a platform a temporal process starts, in which users will rate uploaded pictures, view, comment or share them, and these properties will change over time [22]. Imagine you have some nice looking pictures (e.g. taken with different camera settings, cropping or subjects) and you do not know which of them you should upload on a photo-platform. One option is that you consider friends for helping you in your photo selection, however it depends on their expertise how they will appreciate and rate your photos. This problem formulation leads to the main research question of our paper: We will analyze if it is possible to automatically predict photo liking. For photo-liking estimation we use the like-view rate ($\#likes/\#views$) as the targeted key indicator. Its behavior can be illustrated as follows: When a lot of people who viewed a photo also liked it, this photo is suspected to have a high likability and will also be liked by other users. Furthermore, other applications are possible, e.g., video thumbnail generation [19] or image ranking.

We will use different feature types for liking estimation, that are based on machine learning or new analysis techniques. In general, features based on well-known rules of thumb for photo aesthetics can be extracted using computer vision approaches [3, 7]. In this paper, we will include complementary features, which consider typical photo-sharing aspects and the provided meta knowledge. Our features include comments, social network data, image classifiers, and pre-trained deep learning networks in combination with several machine learning techniques.

Most of all features for aesthetic prediction in the literature only use photo-related aspects. There are some based on social network analysis [21], however they were not combined with other analysis approaches, or are only based on restricted social-media knowledge. One reason for this observation could be that most publicly available image datasets do not include social or meta-data, due to the contained user-specific profile-data. These problems can be circumvented to some extent, if the datasets are stored in an anonymized way.

Our general idea is to use analysis techniques to derive features and combine them with classical image-related features. After deriving and defining all features, we will describe the details of our regression-based approach for prediction of like-view rates.

The effectiveness of our new features will be evaluated in a large-scale experiment using images and meta information from the 500px photo sharing platform. We consider different feature groups (technical, low-level, high-level, social-network-based and deep-learning-based) to show, in different experimental settings, that our new features can improve regression accuracy for like-view rate predictions.

[1] for Flickr: average 1.68 million photos per day for 2016, see https://www.flickr.com/photos/franckmichel/6855169886/

II. Related Work

We first briefly review the literature and features for aesthetic and liking estimation of images. Furthermore, we will analyze previous approaches for image liking prediction and compare them considering ideas for extensions. Important is a

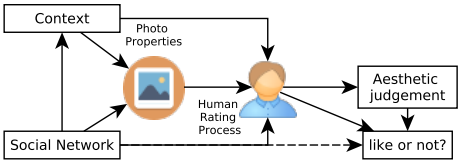

Fig. 1. How humans rate aesthetic and decide liking, based on [10].

proper definition of photo aesthetic appeal based on a generally workable understanding of the term. With the complementary analysis of social network aspects we acknowledge that appeal and liking may be correlated, yet are not the same [18, 5]. Human rating of images and in a social platform liking decisions depends on several influence factors. In Figure 1 a summarized and extended (for the final liking decision) view of Leder et al.'s model of aesthetics ratings is presented. Three main factors (that influence each other) are important for human aesthetics judgment: the photo, the context and the social influences [10]. Finally, based on the aesthetics rating and social impact of a shared photo the human will decide if the photo will be liked or not. Liking prediction is therefore related to aesthetic prediction combined with image appeal (e.g. technical properties of the image) and social network properties of the user (e.g. usage pattern of the user, users community, ...).

A. Principles of Photo Aesthetic/Appeal

Joshi et al. analyze perception of aesthetics as a function of artwork, intent of artist, genre of art, perception, ... and level of experience of the viewer [6]. They differentiate aesthetic in two general terms, 'true aesthetics' and 'observed aesthetics'. For 'true aesthetics' an infinite expertise is required, considering a global view of all images, and 'observed aesthetics' subjectively depends on the observer's attitude. There are several open questions and problems in the field of image aesthetics, for example for artwork characterization, the social impact of aesthetics, or users' emotion prediction. Real-world imaging applications can benefit from an estimated value of aesthetics, such as image retrieval systems or camera guides. The meta information from photo-sharing platforms can be used to provide additional information for analysis of the users' impact on image aesthetics, for example in terms of their role and own expertise [9, 8]. Here already, the lines between aesthetics and liking are starting to get blurred. Considering that image aesthetics are assessed by individual viewers, it is clear that users' experiences and social network have a large impact on perceived image quality. For example, Lebreton et al. analyzed relations between the users' knowledge in photography and their rating behavior in a crowd-sourcing study [9].

B. Feature Definition and Prediction Approaches

Previous research is based on feature sets that were derived using low level image properties, meta-data and community effects. For example, Datta et al. defined low-level visual features such as exposure of light and colorfulness, average saturation and hue, rule-of-thirds-based values, ... [3] They used feature selection and classification or regression approaches for image appeal based on a dataset of about 3500 images. When using all 15 defined features they were able to achieve a classification accuracy of about 70%, in contrast to rather bad results obtained for their regression experiment. Wu et al. used an SVM-based approach with a sigmoidal softening function and distinguished six classes from good to ugly of image appeal [23]. The authors used a dataset containing approximately 11k photos from Flickr and extracted low-level features (similar to [3] combined with some new features). Features are based on the HSV color space (average values for global or central hue, saturation, and value), position of the main object and colorfulness (using color histograms), combined to a 39- dimensional feature vector. In general, the SVM and sigmoidal softening approach is able to perform well (around 0.8 prediction accuracy). The influence of image aesthetics on liking in photo-sharing platforms is based on more than only low-level features. Subject, image composition and social influences are other factors for aesthetics and liking. Khosla et al. extended low-level features with computer vision features [7]. They analyze popularity of photos in terms of log-normalization of view count using a large dataset from Flickr with approximately $400k$ users. For view-count estimation, they use an SVR approach with three feature groups: low-level features (e.g. HSV color space, color histograms, similar to [23]), computer-vision features (e.g. ImageNet classification, color patches) and high-level features (e.g. object recognition, social cues). Best results, with a rank correlation of approximately 0.8, were achieved using all feature groups. Their high-level features are based on the social network part of Flickr, e.g., mean views of all pictures of a user, photo count of a user, number of contacts, title length, and more. These social-network features can be extended, e.g., a typical photo title includes more semantic information. Such a title is a short summary of the photo and what the photographer wants to express. Comparing with photo tags a new feature can be derived that is measuring semantic similarity of title and tags using, e.g., word vector representations calculated using Mikolov et al.'s [13] model.

Most analyzed approaches for image aesthetics do not combine all of our identified different features. Therefore, only subsets of feature groups based on object description, sentiment, technical features, deep learning, computer vision features, or social network effects were already analyzed in the aforementioned literature.

Our approach combines and extends all of these features to estimate liking values. We are able to compare our new generated features with already published feature sets, analyzing the introduced performance gains.

III. Our Approach

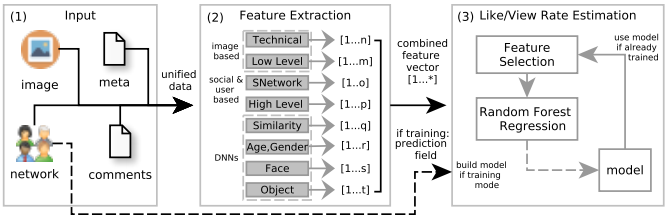

Fig. 2. Steps of our Machine Learning Framework: (1) pictures, additional data were used for extraction of defined features (2) and prediction values for training or using a machine learning algorithm (3).

Our general approach is divided into two phases using three steps each, see Figure 2. To predict liking values we use a machine learning approach implemented with scikit-learn [15]. For each photo we assume that the image itself, meta information (e.g. camera type, exposure) and some social data is available (1). As first phase we train a machine learning algorithm with calculated features (2) for each photo of the training set, so that a model (3) can be derived. For training, it is required that we know the prediction values. The second phase in our approach needs a set of photos with all additional feature-related information, which we use to estimate like/view rate values. In our pipeline, we use feature selection based on an extra tree regression approach to decrease feature space dimensionality, so that only important features will be included in our last step. As final step, we use a random forest regression model. We analyzed and optimized the number of used trees in small experiments and found that 100 decision trees are a good trade-off for speed and prediction performance (all other parameters performed best using default values). Furthermore, we analyzed in small experiments (sample of 3600 images) which prediction variable should be used. Most like/view rates were small numeric values in a small range, so we finally used a modified rate $log(\#likes)/log(\#views)$ as the liking prediction variable to circumvent these problems. These values are in a wider range and less numerically unstable than pure like/view rates. In the next sections we will show all features and describe how they can be calculated.

A. Technical Features

Modern cameras store a lot of meta information when shooting a photo, e.g., camera settings (lens data, aperture, camera model, exposure time) or date and location. We will use all these stored meta information of a given picture. We assume that some technical settings clearly influence the appeal of a photo. For example if you want to smooth water disturbances in a photo of a waterfall, you need to increase the exposure time. For date-information we extract hour of day and week of the year (season indicator) as feature values, e.g., to considerate that photos taken in the noonday sun have a worse quality than photos shot at a different time of day. Not all technical rules can be applied to every picture in every scenario, however they can be used as a starting point. How technical features perform for image liking prediction is one question that we will analyze in our experiments in Section IV.

B. Low-level Features

We define low-level features as quickly calculable features that are based on image analysis or filters. Wu et al. successfully used a number of such low-level features. We will use global hue, saturation and value, and extend central HSV using sub-images based on rule-of-thirds. So for every sub-image (1...9) we calculate mean values for HSV as features as *subimageH/S/V 1...9*. Additionally, we store minimum and maximum indices of each sub-image's HSV values for determination of important or non-important regions of the image. We here assume that the most important image information is located in one of the sub-images, Wu et al. only assumes that this is in the center. Another feature that is based on ideas from [23, 3] is our color-distribution feature. We extended it by using a histogram of up to 1000 distinct color values. A further feature is image noise. Because noise pixels will influence perception of details or filled areas. Image noise mostly depends on the camera sensor and environmental factors. For a coarse estimation, we apply a simple median noise reduction filter on a given image and calculate differences to the original image as *NoiseDiff* feature. Therefore, *NoiseDiff* is an indicator of how much noise was in the given picture, or how many changes were applied by the simple de-noising. *EdgeRatio* works similar to *NoiseDiff*. We apply an edge detector on a given image and calculate the ratio of detected edge-pixels to all pixels. *EdgeRatio* is an indicator of how many edges are present in a given image. Considering that some appealing images have fewer edges or at least the main object is not dominating the whole image.

C. High-level Features

In contrast to low-level features, our high-level features depend on the user. We define high-level features as features that are based on user's knowledge and interaction during the upload, e.g., assigning tags, title or descriptions. They cannot be derived directly based on image analysis. We extend a pure title length feature [7] with natural language processing using nltk [1] to several features. First of all, *titleWordCount* is the total count of tokenized title words, we use a reg-ex based tokenizer for tweets. Second, *titleNonStopWordCount* is the count of tokenized non stop words in a title. It is based on the observation that stop words in titles are mostly just fill words (stop words are typically excluded during natural language processing). We also define *titleWordLenDist* as a histogram of tokenized title word lengths. As further extended features, we calculate *titleTagWordSim* as word vector similarity of title and tags using word2vec [13], *titleTagJaccSim* as Jaccard similarity of tag and title word tokens and *titleSentiments* as sentiment classification values (similar to comment sentiments). We define these similarity features to measure the intention of the title, e.g. is the title related to the subject that the photographer selected. Finally, to sum up, our high-level features are only based on additional information that e.g. a photographer would add in a private photo collection and that are mostly added in photo sharing platforms.

D. Features using pre-trained Deep Neural Networks (DNNs)

We use several pre-trained deep neural networks to generate features for our framework. Using those pre-trained networks for different image tasks (object classification, similarity or face and age detection) we are able to assign each feature a meaning that can be used in a later evaluation to derive liking rules. DNNs perform well for several image analysis problems, which is why we will use a pre-trained image classification model [20] to derive new features. In our framework, two multi-value features will be calculated based on the inception network [20]. First, we predict top-5 classes with associated scores and use these values *classDistScores* as distribution feature. Using *classDistScores* we are able to analyze what content is shown in the picture, to get a deeper understanding of the subject. Our second DNN-feature called *lastLayerValues* uses next-to-last layer of inception as feature vector, consisting of distinct 2048 values. To derive another set of features we use techniques from image retrieval [11]. Based on a deep learning network for calculation of image similarity hashes we conclude two new features, *Hash* and *HashProbs*. The first feature *Hash* is the hash value as integer, because we assume that similar images have similar appeal. For the second feature *HashProbs* we use the provided probabilities (48 float values) of next-to-last layer. There are often times faces and people on images that show high appeal, for example portrait photos. Hence, we use gender, age [16, 17], and face [14] detection networks and derive *Age*, *Gender*, *FaceCount*, and *MaxProbFace* as features. *FaceCount* is based on an estimation of probabilities for pre-trained faces and filtering by a given threshold. For this reason we also define *MaxProbFace* as feature, it serves as indicator for how accurate the feature *FaceCount* is. The DNN-features can be calculated using only the provided images, so that this feature group is independent of meta or social network data and can later be used in a scenario where only the access to the image is possible.

E. Social Network (SN) Features

We define social-network features as features that are based on comments and social network characteristics. For comments we use sentiment classification, because users' comments are able to describe and judge images. A pre-trained sentiment classifier [12] is used to calculate polarity and subjectivity values for a given text. For all comments $C = [c_1, .., c_n]$ individually we extract the following values: median, mean, and variance values of all polarity and subjectivity measurements. Additionally, we combine all comments to one text and calculate global values. In sum, we specify eight features for each picture using sentiment of comments as *commentSentiment*. Furthermore, we calculate median, average, variance, maximum, and minimum of lengths for each comment word using a tokenizer [1] as *commentWordLength*. Based on features introduced in [7] we define *usersAffection*, *followers-*, *friends-*, *galleries-*, *groups-*, *favorites-* and *photosCount* in a similar way, they are provided directly in the 500px platform. We also define the friend's comment rate, assuming that friends' comments may have a large impact in a user's local social network. *friendCommentRate* is defined as ratio of how many friends commented on a photo to all comments. For example, if the local network of a given user is large, many users will be reached after submitting a photo, and it is easier to like photos of friends. We also introduce three new features based on social network analysis of the local network view. Because local friends are more important, we will only analyze a maximum of 3-hop-friends. *triangleCount* is based on counting triangles in a graph and is a reduced form of the centrality metric for social networks [2]. We assume that the local network of a user u is a graph $G_u = (V, E)$ based on extracted values for friends of friends. For each 3-vertex clique of u we count one triangle, assuming that a higher connectivity in a local social network will speed-up commenting and rating of a photo in such a platform. We further define the mean friend count of all friends of a given user u as *MeanFoFCount*. So we calculate average friends count of each second level friend of u (a second level friend of u is a friend of a friend of u). Our last new feature is based on counting how many users can be reached in our local network. It is defined using the number of friends that are reachable in two steps in G_u starting from u and named as *2hopReachableUsers*. *2hopReachableUsers* is an indicator of how many friends are reachable in two steps based on a user u in our local social-network G_u.

TABLE I
FEATURE GROUPS, **S**=STRING, **I**=INTEGER, **F**=FLOAT, **M**=MULTIPLE, SN=SOCIAL NETWORK, LN=LOCAL NETWORK, * INPUT DEPENDENT

Group	Name	Type	Src	Dim
Technical	cameraType, locationName	S	meta	2
	focalLength, ISOValue, shutterSpeed	I	meta	3
	latitude, longitude, aperture	F	meta	3
	height, width, dateInfos	I	meta	4
Low-Level	globalHue/Sat/Val	F	img	3
	subimageHue/Sat/Val 1...9	MF	img	27
	max/min index of SubImgHue/Sat/Val	MI	img	8
	colorDist, noiseDiff, edgeRatio	MF	img	*+2
High-Level	titleWordCount, -NonStopWordCount	I	meta	2
	titleWordLenDist	MI	meta	*
	titleTagJaccSim, -TagWordSim, -Sent	F	meta	3
DNNs	classDistScores, lastLayerValues	MF	img	2053
	hash, hashProbs	I	img	49
	age, gender, faceCount, maxProbFace	F	img	4
SN	commentSentiment, -WordLengthDist	MF	com	6+*
	comment, friendCommentRate	MF	com	*+1
	followers/friends/galleries/groups-Count	I	user	3
	userAffection/photos/favorites-Count	I	user	3
	LN-triangleCount/MeanFoFCount	I	user	2
	LN-2hopReachableUsers	I	user	1

F. Summary of Features

Table I summarizes all presented features detailed. For example, feature *commentSentiment* in feature group *social network* is a multi-value float feature based on comments of a given image. For all defined features we specify which source is required (image, meta data, user info or comments). In summary, we defined 11 technical, 39 low-level, five high-level, 13 social network and eight deep learning feature-sets, whereby a high dimensional feature space is defined (most features have multiple values). To avoid that our final system just uses, e.g. DNN based features, we use a feature selection step (that filters our unimportant features).

IV. Experimental Evaluation

We conducted two experiments for evaluation of our approach. Our first experiment IV-A uses different feature group settings with a small dataset to compare the effectiveness of our added features to that of the low-level features (LLF). The main goal of our first experiment is to evaluate how much our new features will improve regression accuracy. Due to the fact that low-level features were already studied in several other experiments [23, 3] we will use the performance obtained with these features as reference baseline. Also, our LLF features are mostly a pure re-implementation of state-of-the art features. Experiment IV-B will focus on two feature sets, namely ALL and the photo-derivable features (OPD) compared to low-level features and use large datasets for a representative analysis. We downloaded pictures, meta-, and social data from 500px of two categories ("editors" and "fresh") using the provided API.

TABLE II
CRAWLED 500PX IMAGES; USED SAMPLES FOR EXPERIMENTS.

discover-category / sample	# crawled pictures	$log(\#like)/log(\#view)$ rate r $\min r$	$\max r$	$\overline{r}$	$\sigma(r)$
editors	20975	0.43	0.83	0.68	0.05
fresh	59130	0.10	0.86	0.62	0.11
sample9000	9000	0.10	0.82	0.62	0.11
combined20k	20000	0.11	0.83	0.65	0.09
all80k	80105	0.10	0.86	0.63	0.10

Furthermore, in Table II our dataset composition is summarized. In our first experiment we will use a randomly chosen sample from our "fresh" images (sample9000). Sample9000 has similar rate properties as the fresh category, min, max, mean and standard deviation are approximately equal, for visualization and finding first directions a smaller sample is more suitable. In our second experiment we used a combination of 20k randomly selected fresh and editors images as dataset (combined20k), with 10k images per each of the two categories. Furthermore, we use in our second experiment a dataset that is a combination of all images of both categories with approximately 80k images.

A. Feature Groups Evaluation

For comparison of our newly provided features we used our low-level features as reference, because similar low-level feature sets were already used in different other experiments [3, 23]. As dataset we use sample9000. We trained two regression algorithms and performed a 10-fold cross validation. In Figure 3 we compare a trained prediction system with all implemented features (ALL) with a predictor that uses only low-level image features (LLF). It is notable that LLF predicts more frequent values for like/view-rates around 0.6 than ALL and also observed in the actual image data. ALL better fits the original rate than LLF. Furthermore, we calculated RMSE (root mean square error) and R^2 (correlation coefficient of determination) for both experimental settings ALL and LLF. For LLF we obtained an RMSE of 0.108 and a R^2 of 0.072

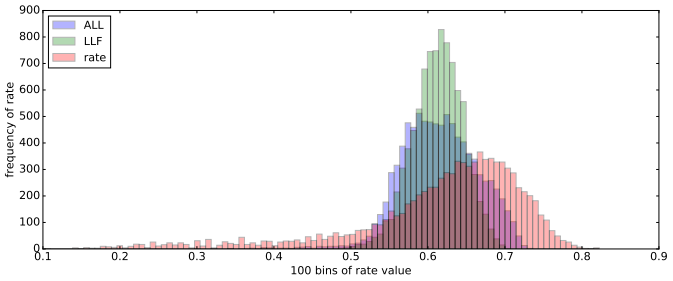

Fig. 3. Distribution of real $log(\#likes)/log(\#views)$ and predicted values of ALL and LFF for sample9000. Note: similiar distributions can include false-predictions.

compared to ALL's RMSE of 0.098 and R^2 of 0.231. Even if ALL is just 10% better comparing the RMSE, the R^2 of LLF is approx 0.1, that means, that there is no linear correlation to the original rates. The R^2 for ALL is better and indicates a (though still low) linear correlation. In contrast to classification experiments from the literature we use a regression model that's why a direct comparison is not possible. We further analyzed the specific impact of our features in this experiment (top-10 important features). The most important features for LLF are *noiseDiff*, *colorDistribution* and *subimageHue 1...9*. For OPD features based on our pre-trained DNNs dominate the important features (*lastLayerValues* and *hashProbs*). Highly interesting results can be observed in the ALL feature set. The most important features in this experiment are social network comment features, photo count and *lastLayerValues* of our DNNs. In our general machine learning pipeline, our first step is selecting important features, hence a small analysis of how many features are important for each feature-set was applied. For LLF, 439 of 1370 features were used in the final prediction. The set ALL uses 2297 out of 7943 distinct features.

Furthermore, it can be observed that each single feature has only a small impact on prediction of like/view rates. Hence, for a better feature evaluation we performed a leave-one-out analysis, based on our defined feature groups (technical, low-level, high-level, social network (SN), and DNNs).

In Table III the results for our leave-one-out approach are summarized. Especially noticeable is that leaving out social network features will significantly decrease the R^2 value, so that our model correlates less well with the original values. This proves our hypothesis that social network features have a high impact on photo liking. In general, all $RMSE$ values are similar, furthermore low-level, technical, and high-level features have about the same influence on R^2.

B. Photo-subject Derivable Features

In this experiment we will further analyze features that are deducible based on a given photo content, referred to as

TABLE III
$RMSE$ AND R^2 FOR LEAVE-ONE-OUT EXPERIMENT; SAMPLE9000.

leave-out-feature	technical	low-level	high-level	SN	DNNs
$RMSE$	0.098	0.098	0.098	0.104	0.097
R^2	0.233	0.228	0.229	0.135	0.255

OPD in comparison to our other feature sets. OPD includes the technical, low-level and deep neural network features. In our feature-set OPD, no social input based on comments or a user's network is required. Thus, this feature-set is important for a pure like-view estimation if for example a user is new in a photo sharing platform. We compare, similar to Experiment IV-A, the performance with our added features to that obtained with the low-level features LLF. We also perform a 10-fold-cross validation for each sub-experiment.

TABLE IV
EVALUATION ALL, OPD AND LLF, * HIGHLIGHTS BEST VALUE.

$RMSE$	ALL	OPD	LLF	R^2	ALL	OPD	LLF
sample9000	**0.098***	0.105	0.108		**0.231***	0.128	0.072
combined20k	**0.077***	0.088	0.091		**0.326***	0.114	0.056
all80k	**0.085***	0.096	0.099		**0.329***	0.139	0.091

In Table IV all results for our 10-fold large experiments are summarized. We calculated, for each feature group ALL, LLF, OPD, and each used dataset the resulting $RMSE$ and R^2 values. For sample9000 ALL performs $\approx 9\%$ better compared to LLF for RMSE , for combined20k we achieved a boost by $\approx 15\%$ RMSE and for all80k the performance is approximately $\approx 14\%$ better . The R^2 values for ALL are always better than for LLF, that means ALL yields a higher linear correlation than LLF. For LLF all R^2 values are approximately zero, so that no correlation occurs. The social and high-level impact of photo-sharing platforms cannot be modeled using LLF.

Comparing to Experiment IV-A, social network features have a similarly high impact in Experiment IV-B. We can conclude that ALL performs as best and OPD can be used for an approximation of image liking if a user just started in the photo sharing platform and has less social connections or impact.

V. CONCLUSION

We introduced a framework for image-liking prediction using an extended set of features. Our features are not only based on pure image information. They include social media and meta information as well as comments. For the first time we combined several analyses and machine learning approaches of social network analysis, natural language processing and deep learning in order to estimate liking values of photos. Our experimental evaluation showed that such a regression approach performs 15% better in terms of $RMSE$ and for R^2 when using our newly introduced features than a system using only low-level features of images. Further, we found that social network features have the largest impact on image liking prediction. Moreover, the dataset we created for this study, based on images, meta information, social network data can be used for several future experiments. Image liking is based on various factors and is hard to predict. So far we only used $log(\#likes)/log(\#views)$ rates for prediction, however there are further indicators for image liking and appeal than like/view rates, which may be influenced by psychological aspects, temporal effects, fashion effects, and art. For our large-scale experiments, those factors could not be excluded, because they are intrinsic to human perception and experience. Further analysis of such influences should be conducted.

VI. REFERENCES

[1] S. Bird. "NLTK: the natural language toolkit". In: *COLING/ACL*. Association for Computational Linguistics. 2006, pp. 69–72.

[2] P. J. Carrington et al. *Models and methods in social network analysis*. Vol. 28. Cambridge university press, 2005.

[3] R. Datta et al. "Studying aesthetics in photographic images using a computational approach". In: *European Conference on Computer Vision*. Springer. 2006, pp. 288–301.

[4] M. Freeman et al. *The Photographer's Eye: Composition and Design for Better Digital Photos*. CRC Press, 2007.

[5] L.-C. Hsieh et al. "Investigating and predicting social and visual image interestingness on social media by crowdsourcing". In: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE. 2014, pp. 4309–4313.

[6] D. Joshi et al. "Aesthetics and Emotions in Images". In: *IEEE Signal Processing Magazine* 28.5 (Sept. 2011), pp. 94–115.

[7] A. Khosla et al. "What makes an image popular?" In: *Proceedings of the 23rd international conference on World wide web*. ACM. 2014, pp. 867–876.

[8] P. Lebreton et al. "Evaluation of aesthetic appeal with regard of user's knowledge". In: *EI* 2016.16 (2016), pp. 1–6.

[9] P. Lebreton et al. "Studying user agreement on aesthetic appeal ratings and its relation with technical knowledge". In: *QoMEX 2016*. IEEE. 2016, pp. 1–6.

[10] H. Leder et al. "A model of aesthetic appreciation and aesthetic judgments". In: *British journal of psychology* 95.4 (2004), pp. 489–508.

[11] K. Lin et al. "Deep learning of binary hash codes for fast image retrieval". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2015, pp. 27–35.

[12] S. Loria. "TextBlob: simplified text processing". In: *Secondary TextBlob: Simplified Text Processing* (2014).

[13] T. Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[14] O. M. Parkhi et al. "Deep face recognition". In: *BMVC*. Vol. 1. 3. 2015, p. 6.

[15] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *JMLR* 12 (2011), pp. 2825–2830.

[16] R. Rothe et al. "DEX: Deep EXpectation of apparent age from a single image". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015, pp. 10–15.

[17] R. Rothe et al. "Some like it hot-visual guidance for preference prediction". In: *arXiv preprint arXiv:1510.07867* (2016).

[18] R. Schifanella et al. "An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures". In: *arXiv preprint arXiv:1505.03358* (2015).

[19] Y. Song et al. *To Click or Not To Click: Automatic Selection of Beautiful Thumbnails from Videos*. 2016.

[20] C. Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". In: *CoRR* abs/1512.00567 (2015).

[21] L. C. Totti et al. "The impact of visual attributes on online image diffusion". In: *WebSci '14*. ACM. 2014, pp. 42–51.

[22] B. Wu et al. "Unfolding Temporal Dynamics: Predicting Social Media Popularity Using Multi-scale Temporal Decomposition". In: *13th AAAI Conf. on Artificial Intelligence*. 2016.

[23] Y. Wu et al. "The good, the bad, and the ugly: Predicting aesthetic image labels". In: *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE. 2010, pp. 1586–1589.