Facial Image Aesthetics Prediction with Visual and Deep CNN Features

Mohamed Selim[1], Tewodros Amberbir Habtegebrial[1], and Didier Stricker[1,2]

[1]Technical University of Kaiserslautern
[2]Augmented Vision, German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
mohamed.selim, tewodros_amberbir.habtegebrial, didier.stricker@dfki.uni-kl.de

Abstract

Large number of images that has persons are being uploaded to the Internet, at a very high rate. However, they vary in quality and aesthetics. These variations affect the performance of the facial images analysis algorithms. This fact poses an interesting question: *Can we predict the aesthetics of the facial image in stills?*. In this work, we introduce a framework that uses deep face representations from CNNs and other visual features to tackle the problem. We evaluated our algorithms on large scale datasets of persons. Regarding the aesthetics, we used collected portraits from the AVA dataset, as well as the Selfie dataset. We thoroughly evaluated our algorithm. Moreover, we outperformed the state-of-the-art in aesthetic prediction in portrait images as we achieved accuracy of 84% while the state-of-the-art achieved 64.25% by using deep representations from our AestheticsNet combined with visual features.

1 Introduction

In this paper we study aesthetics problem of facial images. Aesthetics prediction has been assessed by computing average aesthetic scores given to images by human annotators directly or indirectly. In some datasets like [Redi et al., 2015], aesthetics scores are given by users. In some cases like [Kalayeh et al., 2015], the scores are inferred indirectly from other information like number of views of the image on a social network. When it comes to the computational modelling of aesthetics of images, aesthetics could be predicted through visual features and from other various cues. Especially on the web, images contain additional information in the form of user comments, number of views and likes by other users etc. However, we restricted the scope of our study to predicting aesthetics only from visual cues.

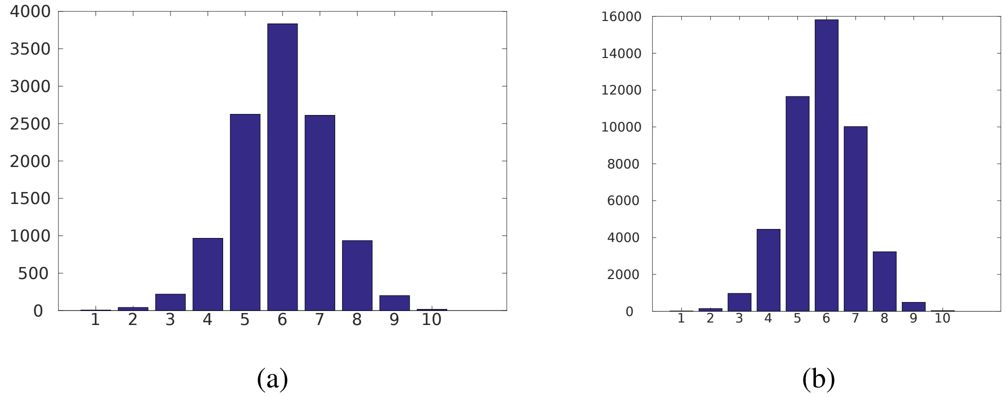
Selfies have become a common phenomenon. Google reported (http://goo.gl/53ZOjG by 2014) more than 93 million self portraits were being taken everyday on android devices. As of April 2016, a keyword search with "#selfie" retrieves 286,297,630 results on Instagram. As a study performed on the popularity of images on social media showed, images with faces are more likely to get comments and likes [Bakhshi et al., 2014]. In our study *"Aesthetics"* of portraits, we used different tools and mechanisms borrowed from computer vision and machine learning. Our image quality prediction algorithm is based on Convolutional Neural Networks (CNNs); for the purpose of image aesthetics prediction we used CNNs and other computer vision features like GIST, HOG and LBPs. We applied these tools on different Datasets of portrait images [Redi et al., 2015], and a collection of selfies [Kalayeh et al., 2015].

2 Related Work

In recent years, a significant amount of work has been devoted to the problem of image aesthetics prediction [Lienhard et al., 2015, Kang et al., 2014]. The works done in the area could be separated into two major groups: *Feature-based* and *Learning-Based*. Feature-based approaches represent methods that predict image aesthetics from low-level visual features and high-level attributes [Marchesotti et al., 2013]. [Dhar et al., 2011] used high-level describable attributes such as presence of people and portrait depiction, object and scene type, etc. However, deep CNNs[Lu et al., 2014] were also used for aesthetics prediction. Learning-based approaches require no hand-crafted features and they are also shown to be more robust. Recently, researchers are focusing on effective aesthetic analysis of facial images. [Redi et al., 2015] studied factors contributing to the aesthetic beauty of an image by studying portraits collected from AVA dataset. Their work uses various features to describe photo composition, quality, emotions, etc. Moreover, due to the recent "Selfie" phenomenon, researchers are studying facial images uploaded to the Internet. A study showed that images with faces get more likes and comments on social networks [Bakhshi et al., 2014]. [Kalayeh et al., 2015] initiated the study on selfies by creating a new selfie Dataset to open the door for more in-depth studies on selfies.

3 Proposed Approach

Datasets For the purpose of aesthetics prediction, we used a newly introduced Dataset of portraits from [Redi et al., 2015]. The portrait dataset contains 11,400 images. Every image in the Dataset has an aesthetics rating value ranging from 0 to 10. These ratings are average ratings given by users of the DPChalenge website. Figure 3 shows the distribution of ratings of the images. We modelled this problem as binary classification problem where labelling all images whose average score was above 5.55 as "good" and those with score below 5.55 as "bad" similarly to[Redi et al., 2015]. As suggested in [Murray et al., 2012, Redi et al., 2015], we also introduced a margin parameter, denoted by parameter δ , for discarding ambiguous images. We discarded all images within 5.55 $\pm\delta$. In our experiments we tested with values of delta 0.1 and 1. The Selfie Dataset [Kalayeh et al., 2015] contains more than 46,000 selfies collected from *selfeed.com*. The popularity score of the selfies was calculated as log_2 *normalized* views counts of the selfies.



(a)

(b)

Figure 1: Score Distribution of (a) Portraits Dataset [Redi et al., 2015], and (b) Selfies Dataset [Kalayeh et al., 2015].

Aesthetics Prediction using Visual Features We analyzed the effectiveness of many computer-vision features in the task of aesthetics evaluation in the context of Portrait images. For this purpose, we extracted features by using GIST (global feature vector for an image by applying oriented Gabor filters at multiple scales), HOG and LBP.

Aesthetics Prediction using CNNs We used CNNs as direct aesthetics prediction by training them on Aesthetics datasets and methods to extract features which are useful for aesthetics. For extracting high level information, we used neural Networks trained on classifying *Image Style* and detecting *Adjective-Noun-Pairs*.

We used different CNN architectures in our work. We used StyleCNN which was introduced to classify image style on Flickr [Karayev et al., 2013]. We also tested Deep SentiBank [Borth et al., 2013] which showed effectiveness in Adjective-Noun-Pairs in visual sentiment detection. We trained AestheticsNet which is based on AlexNet [Krizhevsky et al., 2012] after changing last layer to predict image aesthetics.

4 Evaluation and Results

We conducted experiments on two different tasks: *Aesthetics of Portraits*, and *Popularity of selfies*. In case we fuse features together, we normalize the feature vectors using sigma normalization (early normalization).

Aesthetics Prediction In Table 1 we present our best combination of features, which clearly shows better results than the results of [Redi et al., 2015]. With the exception of the CNN, our tests on the Portraits Dataset [Redi et al., 2015] are done in a 10-fold-cross validation manner. Due to the high computational cost of the CNNs, we experimented by splitting (randomly) the dataset in to 8000 training and 3400 test images.

Delta	SentiBank + Flickr Styles	SentiBank + AestheticsNet	Redi et. al. [Redi et al., 2015]
0.1	66.65	67.32	64.25
1.0	**80.14**	**84.00**	64.25

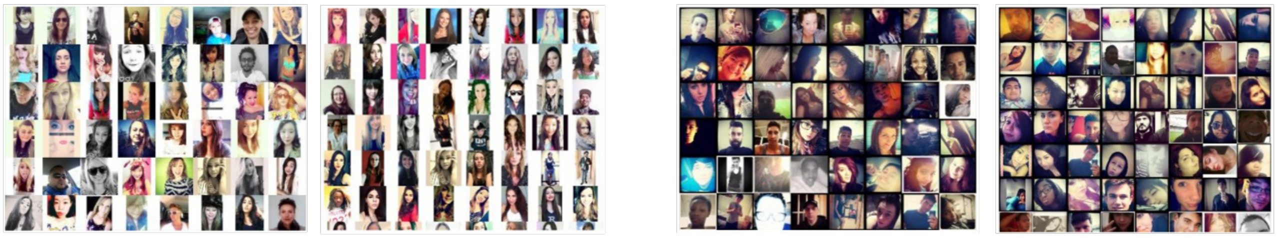
Table 1: Results on combining features for portrait aesthetic classification. We outperform SOA by 2̃0%.

Selfie Popularity Prediction We have used our computer vision features and CNN features in Selfie Popularity prediction task also. As shown in Table 2 the combination of CNN, HOG and GIST gives a 0.59 Spearman's rank correlation, which is better than the 0.55 state-of-the-art (SOA) correlation, reported by Kalayeh et. al. [Kalayeh et al., 2015].

LBP	CNN	Style	SB	HOG	GIST	CNN-HOG-GIST	Kalayeh et. al. [Kalayeh et al., 2015]
0.11	0.45	0.36	0.45	0.49	0.52	**0.59**	0.55

Table 2: Spearman's rank correlation for popularity prediction on Selfies [Kalayeh et al., 2015]. We outperform the SOA.

In the selfie popularity prediction, the GIST feature was very efficient, achieving a 0.52 correlation level. In Figure 2, we show the effectiveness of our algorithms by presenting sample images classified as bad or good. In Figure 2, we show the top images 64 and bottom 64 images in the prediction by the CNN and the ground truth from the selfie dataset. The CNN was able to differentiate between good and bad selfies according to their popularity scores.



Groundtruth: Good

Prediction: Good

Groundtruth: Bad

Prediction: Bad

Figure 2: Sample results from our CNN trained on the Selfies Dataset[Kalayeh et al., 2015]. We show visually that the ground truth and prediction are similar

5 Conclusion

In this paper we propose an approach that combines computer vision features and Convolutional Neural Networks in Facial Image Aesthetics prediction. Our results indicate that using CNNs for learning to detect high level features like computer vision and Image Style can significantly improve the classification accuracy of an aesthetics classifier. In predicting Portraits' aesthetics and selfies' popularity, DeepSentibank's ANPs were better than a CNN trained on the dataset. These results indicate that using a CNN trained on a relevant problem can easily solve other problems. Computer vision features like HOG and GIST were surprisingly effective in predicting popularity of selfies. Thus, by combining various features we created an aesthetics prediction algorithm which outperforms the SOA, as we reach classification accuracy of 84% while the SOA reaches only 64.25%. The results we achieved so far, opens the door for investigating our proposed approach on videos datasets and tackle visual challenges existing in videos compared to still images.

Acknowledgments. This work has been partially funded by the University project Zentrums für Nutzfahrzeugtechnologie (ZNT).

References

[Bakhshi et al., 2014] Bakhshi, S., Shamma, D. A., and Gilbert, E. (2014). Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the SIGCHI*.

[Borth et al., 2013] Borth, D., Ji, R., Chen, T., Breuel, T., and Chang, S.-F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *21st ACM international conference on Multimedia*.

[Dhar et al., 2011] Dhar, S., Ordonez, V., and Berg, T. L. (2011). High level describable attributes for predicting aesthetics and interestingness. In *CVPR*.

[Kalayeh et al., 2015] Kalayeh, M. M., Seifu, M., LaLanne, W., and Shah, M. (2015). How to take a good selfie? In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*.

[Kang et al., 2014] Kang, L., Ye, P., Li, Y., and Doermann, D. (2014). Convolutional neural networks for no-reference image quality assessment. In *CVPR*.

[Karayev et al., 2013] Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., and Winnemoeller, H. (2013). Recognizing image style. *arXiv preprint arXiv:1311.3715*.

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

[Lienhard et al., 2015] Lienhard, A., Ladret, P., and Caplier, A. (2015). Low level features for quality assessment of facial images. In *10th Int. Conf. on computer Vision Theory and Applications, VISAPP*.

[Lu et al., 2014] Lu, X., Lin, Z., Jin, H., Yang, J., and Wang, J. Z. (2014). Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 457–466.

[Marchesotti et al., 2013] Marchesotti, L., Perronnin, F., and Meylan, F. (2013). Learning beautiful (and ugly) attributes. In *BMVC*, volume 7, pages 1–11.

[Murray et al., 2012] Murray, N., Marchesotti, L., and Perronnin, F. (2012). Ava: A large-scale database for aesthetic visual analysis. In *CVPR*.

[Redi et al., 2015] Redi, M., Rasiwasia, N., Aggarwal, G., and Jaimes, A. (2015). The beauty of capturing faces: Rating the quality of digital portraits. In *Automatic Face and Gesture Recognition (FG)*.