Master Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

Human face synthesis for dataset augmentation

Bc. Matěj Nikl

Supervisor: Ing. Vojtěch Franc, Ph.D.
January 2020

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Nikl** Jméno: **Matěj** Osobní číslo: **406558**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra počítačů**

Studijní program: **Otevřená informatika**

Studijní obor: **Umělá inteligence**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Syntéza lidské tváře pro rozšíření trénovací množiny

Název diplomové práce anglicky:

Pokyny pro vypracování:

The current state-of-the-art methods for face recognition are based on deep Convolutional Neural Networks which rely on large databases of annotated facial images. For example, learning predictor of biological age requires large set of examples of facial images along with groundtruth age of captured subjects. Collection of such example sets is expensive. The goal of this thesis will be to develop a method augmenting the limited set of annotated real faces by generating synthetic photo-realistic faces with the same attributes. The success of the method will be evaluated via measuring accuracy improvement of a CNN learned from the generated synthetic images. In particular, the problem of age recognition will be used as a benchmark. The output will be i) a software generating faces of virtual identities with known biological age, ii) a software learning and evaluating a CNN for age prediction and iii) a statistical evaluation showing impact of the virtual faces on the prediction accuracy.

Seznam doporučené literatury:

[1] Wang Y., Liu Z., Guo B. (2011) Face Synthesis. In: Li S., Jain A. (eds) Handbook of Face Recognition. Springer, London
[2] Zhao, J. et al. Dual-Agent GANs for Photorealistic and Identity Preserving Profile Face Synthesis. NIPS 2017.
[3] Franc, V. and Cech, J. Learning CNNs from Weakly Annotated Facial Images. Journal of Image and Vision Computing, 2018.

Jméno a pracoviště vedoucí(ho) diplomové práce:

Ing. Vojtěch Franc, Ph.D., Strojové učení FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **14.02.2019** Termín odevzdání diplomové práce: ______________

Platnost zadání diplomové práce: **20.09.2020**

Ing. Vojtěch Franc, Ph.D.
podpis vedoucí(ho) práce

podpis vedoucí(ho) ústavu/katedry

prof. Ing. Pavel Ripka, CSc.
podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

.
Datum převzetí zadání

Podpis studenta

Acknowledgements

I gratefully thank my supervisor Ing. Vojtěch Franc, Ph.D. for his guidence and help throughout the whole process of working on this thesis. I also thank my girlfriend and my family for their endless patience and support.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prague, January 7, 2020

Matěj Nikl

Abstract

Training a Convolutional Neural Network (CNN) for the task of age prediction from face images requires large amounts of training data. As the freely available datasets are underrepresented for certain age categories, the CNNs cannot learn to recognize them well enough. We propose a solution to this problem by introducing a method for synthesis of new face images – of virtual identities. It creates new examples by merging appearance and shape of several compatible face parts together. Since the face parts always come from faces of the same age and gender, the synthetic example retains these attributes. We evaluate our method using a CNN and on the MORPH dataset. We try to find the best hyper-parameters of our method by performing an exhaustive search. Then, we train and evaluate the CNN in various settings, while measuring its performance using several metrics. Sadly, we observe no significant improvement using our augmentation method. We hypothesize that CNNs are able to learn themselves that the correlation between face parts can be safely ignored for the age recognition task. Hence synthetic examples obtained by permutation of face parts already contained in the training data do not improve generalization of the CNN.

Keywords: dataset augmentation, face synthesis, age prediction, CNN, poisson image editing

Supervisor: Ing. Vojtěch Franc, Ph.D.

Abstrakt

Pro trénování konvoluční neuronové sítě (CNN) pro odhad věku z obrázků tváří je potřeba velké množství trénovacích dat. Protože volně dostupné datové sady mají určité věkové kategorie nedostatečně zastoupené, konvoluční sítě se je nemohou naučit odhadovat dostatečně dobře. Jedním z řešení tohoto problému by mohla být naše metoda pro syntézu nových obrázků tváří – nových identit. Nové příklady tvoří kombinováním vzhledu a tvaru několika částí kompatibilních tváří. Vzhledem k tomu, že takovéto části vždy pochází z tváří stejného věku a pohlaví, vytvořená syntetická tvář si tyto atributy zachovává. Naši metodu hodnotíme pomocí konvoluční neuronové sítě a na datasetu MORPH. Pomocí hrubé síly se pak pokoušíme najít nejvhodnější hyperparametry naší metody. Následně sledujeme hodnoty několika metrik konvoluční sítě trénované a vyhodnocované v různých podmínkách. Naneštěstí se nám nedaří pozorovat významné zlepšení, zapříčiněné naší metodou. Naší hypotézou je, že konvoluční neuronové sítě se v úloze odhadování věku naučí ignorovat korelace mezi částmi obličejů. Z toho důvodu syntetické příklady získané permutací částí tváří z již přítomných trénovacích dat nezlepšují konvoluční síti schopnost zobecňovat na nové identity.

Klíčová slova: rozšíření datové množiny, syntéza tváře, odhad věku, CNN, poisson image editing

Překlad názvu: Syntéza lidské tváře pro rozšíření trénovací množiny

Contents

Figures

Tables

Chapter 1

Introduction

Age prediction is both theoretically interesting and practical problem with increasing number of commercial applications. For example, age-based restriction limiting virtual or even physical access to certain premises, websites, or even denying the purchase of certain goods (such as alcoholic beverages) is a one use-case. Others include customer age estimation in retail and forensic science.

The current state-of-the-art in age prediction is based on CNNs learned from vast amounts of examples. Collecting facial images with annotated age is difficult. A common approach to enlarge the training dataset is based on generating synthetic examples. In most cases, the synthetic examples are obtained by applying a set of geometrical or photometric transforms on the source annotated image. This approach helps to make the trained CNN more robust against changes in pose, rotation, scale and lighting conditions. However, it has a little impact on generalization to unseen identities. It has been observed that adding examples of new identities improves accuracy of face-verification systems more than adding new examples of identities already contained in the training set.

In this work we propose a method to generate synthetic examples of new identities. We try to accomplish this by merging appearance and shape of several indentities together. The whole process consists of several steps. First, we estimate position of the facial landmarks. From these we further estimate other important characteristics, such as the face pose and the (simplified) face expression. Then, we find tuples of compatible faces to be merged together. Finally, while controlling the shape and appearance transfer, we stitch certain face parts together using affine transformations. We use the seamless cloning algorithm [31] as a post-processing step to improve the overall image quality after stitching.

We evaluate the capabilities of the proposed method using a CNN model, which is trained and evaluated in various settings on the MORPH dataset [32]. While we verify that the generated examples are indeed different from the original ones and mostly indistinguishable from real faces, they unfortunately do not seem to provide any new useful information for the CNN to train from. Hence, we observe no significant accuracy improvement when the training dataset is extended by the synthetic examples. We hypothesize that CNNs are able to learn that for the age recognition task, the correlation of face parts can be safely ignored. Therefore, synthetic examples obtained by permutation of face parts that have been already contained in the training set bring no new information for learning the CNN.

This thesis is structured into five chapters. Chapter 2 introduces related methods both for the age prediction task, and for the general and face-specific dataset augmentation. It also gives further motivation for this work. Chapter 3 introduces our proposed method in detail. Chapter 4 explains all the details and results of the conducted experiments. Finally, Chapter 5 concludes this thesis.

Chapter 2

Related methods

Age estimation is a long standing and still unsolved problem. Before the deep learning revolution occurred, methods based on local appearance descriptors such as HOG [7], texture descriptors such as LBP [28] and SURF [3], and others were relatively successfully applied. For example [19] examines the usage of the mentioned features together before applying the Canonical Correlation Analysis (CCA) [18] for the final age estimation. This approach yields the MAE of 4.25 on the MORPH dataset.

As deep learning became more widespread, convolutional neural networks found their way into the age estimation problem. They were useful even as a mere feature extractor for the final age predictor [36, 13], e.g. achieving MAE of 4.77 on the MORPH dataset.

Further works usually employ the end-to-end training regime of CNNs, i.e. training the CNN to predict the age directly (using some encoding) [23, 10, 19, 24, 2, 1, 27, 22, 34, 29], achieving MAE on the MORPH dataset as low as 1.96 [15]. The main differences between the referenced works consist of:

- the input structure and its preprocessing (e.g. multiple aligned patches of face as the input, registration, etc.),
- the architecture of the model, and
- the output age encoding and the training loss function.

2.1 Labeled data scarcity

One of the reasons why the deep learning approach hasn't improved upon the previous results instantly is the scarcity of labeled training data. Moreover, the labels are often noisy and heavily unevenly distributed. Figures 2.1 and 2.2 show the age distributions for various available datasets, and Table 2.1 gives a summary of age range present, the overall example counts and a measure of their non-uniformness.

The target ages are usually either manually assigned based on the appearance (so called the apparent age) as is the case for the APPA-REAL and ChaLearnAge datasets, or automatically assigned e.g. based on the known

Dataset name	age range	# examples	$D_{KL}(\text{pmf} \parallel U\{0;100\})$
AgeDB [26]	1–100	16488	0.387
APPA-REAL [11]	1–100	7591	0.491
UTKFace [37]	1–100	23468	0.552
IMDB-WIKI [33]	0–100	520717	0.641
CACD2000 [6]	14–62	163446	0.810
ChaLearnAge [12]	2–87	3611	0.837
FG-NET [21]	0–69	978	0.900
MORPH [32]	16–77	55095	0.933
total	0–100	791394	0.628

Table 2.1: Summary of available datasets for age estimation, sorted by their non-uniformness

date of birth and the date of capture of the photograph (IMDB-WIKI and UTKFace datasets). In case of the automatic procedure, the error comes e.g. from incorrect identity assignment.

Figure 2.1: The histogram of various datasets for the age estimation task

2.2 A survey of methods used to generate synthetic examples

In the supervised learning setting, the dataset augmentation is a method of expanding the training dataset by creating new training examples with known label. When the input data happen to be images, a whole list of augmentation methods becomes available for use. Such methods simply transform the input image into a different one while preserving the label.

Modern deep learning frameworks such as PyTorch[30] often provide a way of performing such augmentations on the fly, even in parallel with the