A Large-Scale Database of Images and Captions for Automatic Face Naming

Mert Özcan[1]
oezcanm@student.ethz.ch
Luo Jie[23]
http://www.luojie.me
Vittorio Ferrari[1]
http://www.vision.ee.ethz.ch/~vferrari/
Barbara Caputo[3]
http://www.idiap.ch/~bcaputo/

[1] ETH Zurich
Zurich, Switzerland
[2] EPF Lausanne
Lausanne, Switzerland
[3] Idiap Research Institute
Martigny, Switzerland

Learning from images and text. Several previous works learn visual classifiers from images and accompanying text captions downloaded from the Internet. Many works focused on associating names in the captions to the faces of people in news images [1, 2, 4], and use this correspondence to train visual face classifiers. This task is challenging due to the correspondence ambiguity problem: multiple persons might appear in an image, and multiple names in a caption. Moreover, persons in the image are not always mentioned in the caption, and vice-versa.

The FAN-Large database. Several image-caption datasets have been released to study the above problems, but they are collected in rather controlled settings, which introduces unwanted biases. For example, the very popular Yahoo News dataset [1, 3] is one of the largest such datasets. It contains 31k detected faces in 20k images. This dataset is relatively easy because in news photos the key persons usually face the camera and occupy most of the image. Moreover, the captions are not fully representative of those that can be obtained from the wild Web, since they all come from a single clean source (e.g. all editors of Yahoo News tend to write captions in a similar style). We believe that a large scale, realistic dataset would be a very valuable resource for studying weakly supervised learning algorithms.

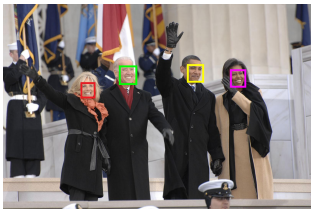
The main contribution of this paper is a very large database of image and captions, coined the *Face and names large scale database* (FAN-Large). The dataset is collected from a wide variety of internet sites. With its 194k detected faces in 125k images, it is much larger than Yahoo News, and the largest dataset of names and faces we are aware of. Every image has an associated natural text caption, extracted from the html page where the image was embedded.

We annotated the names of all the faces in the dataset as well as the verbs indicating the action performed by the persons, when visible. Since the images of FAN-Large are collected from the whole web, its noise level is much higher compared to images collected from a controlled news sources, i.e. the percentage of names which do not appear in the image. The dataset contains 35k unique names and 10k unique verbs. There are 1.4k names that appear 20 or more times, which enables to perform large scale face learning and recognition experiments. Moreover, most captions also contain other types of words than names, such as verbs and adjectives, which enables to study the joint modeling of different type of words [4]. We also systematically collected meta-data such as url and html tags, which might also provide context to facilitate learning.

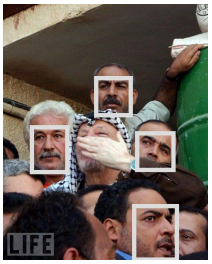
The size and rich information content of this dataset allows us to perform different kind of experiments to study the behavior of weakly supervised learning algorithms. Specifically, beside using the whole database (All), we also consider several interesting subsets with different characteristics (*e.g.* noise level, size of the faces, source websites). For example, the Hard subset contains 26k items with 3 or more names in the caption, while the Buddies subset contains items with people frequently appearing together. Statistics of the whole database as well as some subsets can be found in the table below.

Dataset	images	faces/image	names/caption	verbs/caption
All	125,479	1.55	1.95	0.81
Easy	39,987	1.19	1.34	0.62
Hard	25,607	1.82	4.19	1.33
Life	17,459	1.38	1.78	1.03
Buddies	13,651	1.68	3.12	1.02
Yahoo News[3]	20,071	1.55	1.49	N/A

Contextual features. As a second contribution we propose contextual features that can be extracted from the caption. These help assessing how likely a name is to appear in the image. Among our contextual features there is the position of the name wrt other names in the caption, the position of the sentence in which the name appears, and Part of Speech (POS) tags (noun, verb, adjective, adverb, preposition or other) of the words in the neighborhood of the name.



Caption: **Jill Biden**, Vice President-elect **Joe Biden**, President-elect **Barack Obama**, and **Michelle Obama** wave to the crowd gathered at the Lincoln Memorial on the National Mall in Washington, D.C., Jan. 18, during the inaugural opening ceremonies. More than 5,000 men and women in uniform are providing military ceremonial support to the presidential inauguration, a tradition dating back to **George Washington**'s 1789 inauguration. (photo by U.S. Navy Petty Officer 2nd Class **George Trian**)



Caption: RAMALLAH, WEST BANK - SEPTEMBER 29: Palestinian leader **Yasser Arafat** gestures supporters with a kiss outside his office as Israeli soldiers lift the siege on his compound September 29, 2002 in the West Bank town of Ramallah. After a personal message from U.S. President **George W. Bush**, Israeli Prime Minister **Ariel Sharon** ordered tanks out of Arafat's headquarters today after a 10-day siege. Israel is still calling for the handover of Palestinian militants suspected to be inside the compound.

We propose an extension of the Graph-based Clustering algorithm of [2] to take into account the contextual features, instead of assuming that every name in the caption is equally likely to appear in the image.

Experiments. The third contribution is a thorough assessment of several algorithms on FAN-Large, including constrained GMM (C. GMM) [1], graph-based clustering (GBC) [2] and constrained K-means (C. K-means). We perform experiments on the whole dataset, as well as on subsets constructed to study the impact on performance of specific dataset characteristics. The results bring interesting insights on the behavior of existing approaches. We also experimentally validated that our proposed contextual features considerably increases the name-face assignment performance of [2], resulting in overall higher performance than all compared approaches (GBC+CF). Interestingly, the improvement brought by the contextual features is greater for items downloaded from the professionally edited life.com website. The percentage accuracy for different algorithms and subsets is shown in the table below.

Method	Random	C. K-Means	C. GMM	GBC	GBC+CF
All	39.4%	42.0%	**48.1%**	47.8%	**50.2%**
Easy	42.2%	54.3%	55.0%	**56.7%**	**58.3%**
Hard	26.2%	22.5%	**31.4%**	29.9%	**31.9%**
Life	36.2%	51.5%	**50.9%**	50.1%	**55.1%**
Buddies	26.9%	33.3%	32.9%	**36.0%**	**41.5%**

Database release. We release FAN-Large at `http://www.vision.ee.ethz.ch/~calvin/fanlarge/`, including all annotations, the benchmark protocol, and the contextual features.

[1] T. Berg, A. Berg, J. Edwards, and D. Forsyth. Who's in the picture. In *NIPS*, 2004.

[2] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Automatic face naming with caption-based supervision. In *CVPR*, 2008.

[3] M. Guillaumin, J. Verbeek, and C. Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *ECCV*, 2010.

[4] L. Jie, Barbara Caputo, and Vittorio Ferrari. Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation. In *NIPS*, 2009.