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

Abstract

We present a large scale database of images and captions, designed for supporting research on how to use captioned images from the Web for training visual classifiers. It consists of more than 125,000 images of celebrities from different fields downloaded from the Web. Each image is associated to its original text caption, extracted from the html page the image comes from. We coin it FAN-Large, for *Face And Names Large scale database*. Its size and deliberate high level of noise makes it to our knowledge the largest and most realistic database supporting this type of research. The dataset and its annotations are publicly available and can be obtained from http://www.vision.ee.ethz.ch/~calvin/fanlarge/. We report results on a thorough assessment of FAN-Large using several existing approaches for name-face association, and present and evaluate new contextual features derived from the caption. Our findings provide important cues on the strengths and limitations of existing approaches.

1 Introduction

A huge amount of images with accompanying text captions are available on the Internet. This motivates the recent interest in using captioned images for training visual classifiers. Exploiting the latent associations between images and text can lead to a virtually infinite source of training annotations, without any explicit manual intervention. Previous works have focused on associating names [4, 12] and verbs [17] in the captions to the faces and body poses of people in news images, on learning scene classification models from tagged photos [8, 21], and on learning object recognition models from an online nature encyclopedia [20]. These tasks are more challenging than standard supervised learning due to the correspondence ambiguity problem: it is not known beforehand which part of the image corresponds to which part of the caption. Moreover, not everything mentioned in the caption appears in the image, and, vice-versa, not everything in the image is mentioned by the caption. Several datasets of images and captions have been released to study the above problems, most of which are collected in rather controlled settings (sec. 2). This raises doubts if results obtained on these

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.

Figure 1: *Example from FAN-Large. The detected faces are colored as the corresponding names in the caption. Blue names do not appear in the image, and gray faces are not mentioned in the caption.*

datasets are indicative of the performance achievable on the wild Web, where the amount of data and the level of noise are much greater.

We believe that a large scale realistic dataset is an essential resource for studying weakly supervised learning algorithms. The first contribution of this paper is a large scale database of image and captions for studying the automatic face naming problem. We call it *Face and names large scale database* (*FAN-Large*, sec. 3). It contains 125,479 images of celebrities from different topics downloaded from the Internet. Every image has an associated natural text caption, which we extracted from the html page where the image was embedded. Moreover, most captions in the database contain also other types of words than names (e.g. verbs, adjectives). This enables to study the joint modeling of different type of words. We annotated using Amazon Mechanical Turk (MT) the names of all the faces in the dataset as well as action verbs correspond to the body poses (if they are visible in the image). The annotated FAN-Large database, the software tools developed for its acquisition, the benchmark protocol and the contextual features can be downloaded freely from `http://www.vision.ee.ethz.ch/~calvin/fanlarge/`.

The second contribution of this paper is a thorough assessment of several state-of-the-art algorithms on FAN-Large (sec. 4). We perform experiments on the whole dataset, as well as on subsets constructed to study the impact on performance of specific dataset characteristics (e.g. level of noise, size of the faces, source websites; sec. 2). The results obtained provide interesting insights on the strengths and weaknesses of existing approaches.

The third contribution of this paper is a set of contextual features that can be extracted from the caption. We use information on word position, name position, sentence position, position indicator tags and part-of-speech tag context to evaluate how likely it is that a name from the caption appears in the image. This information is then incorporated as a prior into the Graph-based Clustering algorithm of [12], and its impact demonstrated experimentally.

2 Related works and datasets

Learning from weakly labeled data consists of building visual recognition models from loosely or ambiguously annotated data. In this setup, each image is assumed to contain one or many regions of interest. In contrast to "strongly" labeled data, where hard labels are available for every region, here the supervised information is "weak" because the labels are only provided at the image-level but are not assigned to the regions. Moreover, sometimes

noise labels may also be present in the weak label set, i.e. labels not corresponding to any region in the image.

The task of learning visual models from images and videos with accompanying captions can be naturally casted into this framework. With the avalanche of available images and videos resources on the web, training data can be obtained at a very low cost. Inspired from this idea, in the last few years many researchers have studied how to exploit different text sources, including scenes images from photo sharing websites with tags [8, 21], news photos with captions [4, 13], and movies with scripts [5, 9]. Among them, images with captions are especially interesting, as they contain richer information. For example, the captions may describe the location and properties of objects, and even the relations between them. Berg *et al*. [4] and Guillaumin *et al*. [13] show that names extracted from news captions using natural language processing (NLP) can be used to cluster faces appearing in news images. Gupta and Davis [14] model prepositions in addition to nouns (*e.g*. 'bear in water', 'car on street'). Jie *et al*. [17] show the benefit of modeling names and verbs jointly for annotating faces and body poses in news images. Wang *et al*. [20] learn visual models of butterfly species from descriptions in an online nature encyclopedia. Farhadi *et al*. [10] present a system which generates sentences describing input images, after training from images with captions.

Using the captions for supervision can also be challenging, since a caption may describe the image only partially, or not at all. Hence, the level of *noise* and *ambiguous* labels contained in captions can be much higher compared to other similar problems. This calls for methods able to perform learning in a highly noisy environment. Moreover, in [5], Cour *et al*. show that the difficulty of learning depends on how often the true label and an extra noise label *co-occur* with each other. Finally, the nature of this problem demands algorithms able to handle *large scale* datasets efficiently: given a large amount of training data, weakly supervised algorithms could outperform supervised algorithms trained from smaller fully labelled datasets.

Various approaches have been proposed to conquer the above challenges: generative models [4, 17], discriminative models [5, 16, 21], and graph-based approaches [12] (see [11] for an in-depth literature review). To study the problem properly, we believe it is crucial to have a benchmark dataset containing all the above challenges. Most recent efforts in the computer vision community have gone in constructing large scale object class datasets, such as ImageNet [7] and Tiny Images [19]. However, there are only few publicly available datasets of image and captions. None of them cover all these challenges. The *Yahoo! News*[1] dataset of Berg *et al*. [4] was first introduced for studying the problem of automatically naming faces in news images. It consists of news images and captions collected from the Yahoo! News website in 2002 and 2003. Unfortunately ground-truth labels are not available with the original version of this dataset. Recently, Guillaumin *et al*. [13] extended it and produced a complete ground-truth annotated version, called *Labelled Yahoo! News*[2]. The dataset contains 31,147 detected faces of 5,873 different people in 20,071 images. To the best of our knowledge, it is one of the largest Image and Captions databases for computer vision research. This dataset is rather easy since in news photos the key persons usually face the camera and occupy most of the images. Moreover, the image-caption pairs are not truly representative of those that can be obtained from the wild Web, since they all come from a single source (e.g. all editors tend to write captions in a similar style). Another

[1] Available at http://tamaraberg.com/faceDataset/
[2] Available at http://lear.inrialpes.fr/data

dataset which was collected to study a similar problem is the *Idiap/ETHZ Faces & Poses* [3] dataset. The images were collected by querying Google-images using a list of keywords and retrieving also a snippet of text as an initial caption. Then external annotators were asked to complete these snippets into realistic captions describing the images. Although the images are significantly harder than Yahoo! News, this dataset is rather small (1703 images). The *IAPR-TC12*[4] contains 20,000 natural images taken from locations around the world, and it was initially released for cross-lingual retrieval. Each image is associated with a text caption in up to three different languages. However these captions were written in a very controlled simplified format, and only describe what appears in the image (i.e. they do not contain much noise, as opposed to real captions from the web).

3 The FAN-Large database

Collecting the data. Motivated by the fact that there is no public image-captions dataset representative of all the challenges of the web, we created a new one. We focus on photos of people, such as news photography and portraits, which allowed us to get abundant images with realistic captions mentioning person names. The images were collected by querying Google-images using a predefined list of keywords, using a modified version of the image crawler of Schroff *et al*. [18]. Following [17], we did not only collect images with faces (using celebrity names as keywords, *e.g*. "Barack Obama"), but also collected images with people performing different actions by querying for action verbs. In practice we used a combination of names and verbs, *e.g*. "Barack Obama" + "shake hands". The database contains images of 448 celebrities from 9 different topics (baseball, basketball, boxing, entertainment, football, ice skating, golf, politics and tennis), and 27 action verbs corresponding to distinct upper body poses (*e.g*. shake hands, swing and hold trophy). In total, we have 2,118 unique queries formed by combinations of different names and verbs. While collecting the images, we defined the minimum size to 400 by 300 pixels, and excluded hand drawings using the interface provided by Google. In total, we collected 247,374 images, as well as the htmls of the webpages where the images were embedded. Using our html parser, we detected 210,255 images with a corresponding caption (image-caption pairs, called *items*). We used two off-the-shelf face detectors [1, 3] and a named entity detector from NLP [2] to filter out items without a face in the image and a name in the caption. After performing the filtering, we retained 125,479 items, for a total of 194,046 detected faces and 244,745 names.

Statistics. We list here few properties of FAN-Large:

- *Noise.* FAN-Large contains images collected from the whole web using the Google-image search engine. We randomly sampled 1K samples gathered from daylife.com and life.com, both of which are large online collection of professional photography. Most of these images have captions written by professional editors. About 76% of the names in the captions have a corresponding face in the image, and 59% of the verbs correspond to actions in the image. For data gathered from other websites these numbers drop to 68% for names and 25% for verbs (estimated from 1K randomly sampled images). These numbers also support our claim that data collected from a specific news site are not representative of the harder challenges in datasets gathered from arbitrary websites. There are more than 25,607 items containing at least 3 names in the caption, which allows us to form challenging subsets to study face naming algorithms (see below).

[3] Available at http://www.vision.ee.ethz.ch/~calvin/faces+poses/
[4] Available at http://www.imageclef.org/photodata

- *Number of classes.* FAN-Large contains 34,645 unique names and 9,990 unique verbs. 1,437 names and 1,716 verbs appear at least 20 times, which also make FAN-Large suitable for evaluating face and pose association algorithms [17].
- *Diversity.* Most of the captions contain also other types of words than names and verbs, that can give information about the location and the scene. For example: "red carpet", "press conference" and "airport" occur 2,075, 1,828 and 577 times, respectively. Since the urls the images were downloaded from and their original html files are stored, they may also contain contextual priors useful to facilitate the learning.

Annotating the database with Amazon Mechanical Turk. We annotated the database using Amazon Mechanical Turk (MT). In total, 2,355 annotators participated, 366 of whom have annotated more than 100 images. We presented an annotator with a random image with a bounding-box around a random detected face. We then asked the annotators to choose from a list of names and action verbs, to indicate "who is the person" and "what is the person doing", or choose "none of the above" if the person's face and action do not correspond to any of the listed ones. The names in the list are those detected in the caption by the named entity detector, while the verbs come from a manually defined list for each topic (*e.g.*, 'hit backhand', 'hit forehand', 'serve', 'hold an object', 'celebrate' for tennis). The annotators were encouraged to write the correct name and/or verb if they chose "none of the above" (this allows to have more complete ground-truth annotation, beyond what is mentioned in the caption). To control the quality of the annotations as well as to unmask malicious annotators, we first defined a random subset of 10K images and had multiple annotators independently label them. The labels of these images were then obtained by majority voting. For each new annotator, we randomly selected a few images from this verification subset and gave them to the annotator. If an annotator's performance on those images is poor, we reject all of her annotations.

Interesting subsets. 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 conducting experiments on the whole database, we considered the following subsets:

- *Easy.* In this subset, every face is larger than 60x70 pixels, and every caption contains at most two names.
- *Hard.* Items with 3 or more names in the caption.
- *Life.* Items collected from `www.life.com`. The images are usually of very high quality, and have accompanying captions written by professional editors.
- *Buddies.* Items with people frequently appearing together. Names that appear at least 50 times together in the whole database are considered *buddies*, *e.g.*, Barack Obama and Hillary Clinton (164 names in total). We selected items with at least two names from this buddies list in their caption.

Table 1 compares the number of images and other statistics over different subsets. We plan to release the whole dataset with the MT annotations, as well the the urls where the images were downloaded from and their original html files. Although this paper focuses on automatic face naming, we believe that this dataset would be very useful for studying other related problems. For example, it could also be used to study action verbs associated to body poses, as in [17].

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 [13]	20,071	1.55	1.49	N/A

Table 1: *Statistics of the whole database and some interesting subsets. The columns are the number of images, the average number of detected faces per image, the average number of detected names per caption, the average number of pre-defined verbs per caption. As a reference, we give the statistics also for the popular Yahoo! News dataset.*

4 Name-face association algorithms

In this section, we list the algorithms used in our experiments. Most existing approaches [4, 5, 12, 17] can be seen as a type of constrained clustering, where each cluster of faces correspond to one name. We consider the following constraints, which are widely used in the literature: (a) a face can only be assigned to a name appearing in its associated caption, or to *null* if it corresponds to none of them, (b) a name can be assigned to at most one face, (c) a face can be assigned to at most one name. For an item with F faces and M names, the number of admissible assignments respecting these constraints is $\sum_{j=0}^{min(F,M)} \binom{F}{j}\binom{M}{j}$. We denote the set of all admissible assignments by $\boldsymbol{A}$.

Random Assignment. This is a simple baseline which does not use any image information: randomly choose an assignment from $\boldsymbol{A}$ according to an uniform distribution.

Constrained GMM [4, 12]. This constrained mixture model approach was proposed in [12]. It is a simplified version of the generative model of [4], which treats captions as bags of names, disregarding potential contextual cues. The model associates a Gaussian density in the appearance feature space to each name as well as to the *null* label. Given an admissible assignment $\boldsymbol{a}$, the likelihood of an item containing F faces is $p(f_1, f_2, \cdots, f_F|\boldsymbol{a}) = \prod_{i=1}^{F} \mathcal{N}(f_i; \mu_k, \Sigma_k)$, where k is the index of the name of f_i given by the assignment $\boldsymbol{a}$, and μ_k and Σ_k are the mean and covariance of the Gaussian distribution corresponding to the k-th name. Different from the classical GMM clustering which maximizes the log-likelihood of all the faces, this approach maximizes the sum over the log-likelihood of all items among the assignment $\boldsymbol{A}$. Thus, we use a generalized EM procedure for this maximization. In the E-step, the algorithm finds the best assignments subject to the constraints, then in the M-step it updates the parameters μ_k and Σ_k given the assignments. In our experiments, we constrained each Σ_k to be diagonal.

Graph-based Clustering [12]. This approach first constructs a similarity graph in which nodes correspond to faces, edges connect faces are weighted by the similarity between two faces. Next, the algorithm searches for dense subgraphs of this graph, with each subgraph corresponding to a name. Thus, the objective function can be written as a maximization of the sum of edge weights within each subgraph, subject to the admissibility constraints above. A local maximum of this objective function is found using an iterative approach.

Constrained K-means. We propose here a simpler method than [4, 12]. All the faces in the dataset are clustered into K clusters using K-means, where K is the number of unique names detected over all captions. We then associate each face to all the names detected in its corresponding caption, and assign a name to each cluster by majority voting over all the names associated to the faces it contains. Finally, a face is assigned the name of the closest

Dataset	Yahoo! News	FAN-Large (All)	FAN-Large (life.com)
'Everyone Appear'	71.8	42.4	55.0
Contextual Features	85.8	70.0	80.4

Table 2: *Accuracy (%) of prediction of a name's appearance using the contextual features on both Labeled Yahoo! and FAN-Large. 'Everyone Appear' is a simple baseline which always predicts that a name appears in the image.*

cluster center among those that have a name appearing in the caption of the item that face comes from. Unlike the other approaches, this method does not respect constraints (b) & (c).

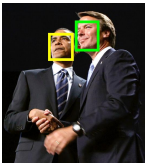
5 The caption contextual features

The algorithms described in section 4 only use the names detected in the captions as supervision. However, captions usually contain richer information than a mere bag of names. Humans can usually guess the content of an image just by reading its caption. For example the "key" persons of the caption usually appear in the image, and the caption contains implicit cues about that. In the computer vision community, some simple contextual cues have been proposed, and they have shown to improve recognition performance [4, 15]. Following a similar spirit, we propose here a larger and more complex array of caption contextual cues for determining how likely it is that a name in a caption appears in the image. We use these contextual features to generate priors, and then incorporate them into a modified version of the Graph-based Clustering algorithm [12].

Cues. We propose the following contextual cues:

- *Word position*: The relative position of the name in the whole caption (i.e. normalized by the number of words in it). This results in a 1-dimensional feature with value between 0 and 1.
- *Name position*: The relative position of the name among the other names in the caption, which results in a 1-dimensional feature between 0 and 1.
- *Sentence position*: The relative position of the sentence in which the name appears (i.e. normalized by the number of sentences in the caption). This results in a 1-dimensional feature between 0 and 1.
- *Position indicator*: When available, position tags behind the name indicate the position of a person in the image, such as "left/(L)", "center/(C)" and "right/(R)". Although these tags are present only in a small fraction of the captions, they are a strong sign that the person appears in the image. We encode a 3-dimensional binary feature, with each dimension corresponding to a position among left, right, center.
- *POS Tags*: We capture here the Part of Speech (POS) tags of the words in the neighborhood of the name, as determined by a language parser [2] (3 words before and 3 after the name). We consider 5 type of tags explicitly (noun, verb, adjective, adverb, preposition), and group all other tags in a 6-th type. For each of the 6 words in the neighborhood of a name, we use a 6-dimensional binary feature indicating its POS tag, resulting in a 36 dimensional feature. This is the most complex cue we propose.

Validation. To validate our assumptions, we performed some experiments on both the Labelled Yahoo! dataset and FAN-Large. We used the proposed contextual cues to predict if a name in the caption appears in the image, without using any image content. Since ground-truth labels are available for both datasets, the task becomes a supervised binary classification problem – does the name appear in the image (+1) or not (-1). We randomly divided both



Caption: GRAND RAPIDS, MI- MAY 14: Democratic presidential hopeful Sen. **Barack Obama** (D-IL) and **John Edwards** (D) shake hands during a rally at the Van Andel Arena on May 14, 2008 in Grand Rapids, Michigan. Former U.S. Senator Edwards announces his endorsement of Obama after Sen. **Hillary Clinton** (D-NY) won the West Virginia primary.

Figure 2: *Example image where contextual features help improve performance. Blue names are assigned to null.*

datasets into training (40% of the captions), validation (10%) and test sets (50%). Then we trained a Support Vector Machine (SVM) [6] on the above caption cues from the training set, and performed feature selection on the validation set using a greedy forward mechanism to discard useless, redundant or contradictory features. This selected 8 features to be used out of the 42 features. The classification accuracy obtained on the test set is reported in table 2. It can be observed that our contextual features improve significantly over assuming that every name in the caption appears in the image. It is also interesting to see that the contextual features achieve higher performance on the Yahoo! News and the Life.com subset than on the whole FAN-Large. A possible explanation is that captions from the same (news) source may be written following a predefined editorial template, which makes it easier to determine if a name appears in the image.

Extended Graph-based Clustering. We propose an extension of the Graph-based Clustering [12] to take into account the contextual features, instead of assuming that every name in the caption is equally likely to be assigned to a face in the image. The original algorithm uses a hyper-parameter which defines the prevalence of *null* assignments. This hyper-parameter acts as a threshold to determine how likely it is that any name will be assigned to *null*, which corresponds to predicting that the name does not appear in the image. The lower the threshold is, the more likely a name will be assigned to *null*. In [12], the threshold is set to a predefined fixed value. We extend the algorithm to vary this parameter according to the output of the classifier based on caption contextual features. If it predicts that a name appears in the image with high confidence, we increase the value of the parameter by a value proportional to the confidence. Hence, the extended algorithm has a null assignment preference parameter *specific to each name*, which varies according to how likely it is for that name to appear in the image according to its caption contextual features. The impact of this extension on name-face assignment performance is reported in the next section.

6 Experiments

In this section we compare the performance of several name-face association algorithms on our FAN-Large database. We used the methods proposed in [9] to extract face descriptors. For the Constrained GMM and Constrained K-means methods, we reduce the dimensionality of the face descriptors to 100 using PCA. We measure performance with two different measures. The first is accuracy: the percentage of correct assignments over all detected faces (including *null* assignments). The second is precision: the percentage of correct assignments over all faces assigned to a name (i.e. not to *null*). To set the *null* assignment preference hyper-parameter of the Graph-based Clustering (GBC) algorithm, we used the heuristic introduced in [12], which assumes that the percentage of *null* faces is known in advance. More precisely, we set this hyperparameter so that the number of *null* faces assigned by GBC is close to the number of ground-truth *null* assignments over the whole dataset. For the other algorithms, there are no parameters to set. Because the name-face association task is essentially constrained clustering, we perform experiments by inputting the entire dataset to

Method	Random	C. K-Means	C. GMM	GBC	GBC (half)	GBC+CF (half)
All	39.4/38.2	42.0/41.4	**48.1/46.2**	47.4/44.4	47.8/45.0	**50.2/51.8**
Easy	42.2/51.3	54.3/56.0	55.0/**59.2**	**56.6**/58.2	56.7/59.3	**58.3/61.6**
Hard	26.2/22.1	22.5/22.3	**31.4**/25.6	28.9/**26.2**	29.9/27.0	**31.9/29.5**
Life	36.2/47.3	51.5/53.1	50.9/52.6	**51.6/53.4**	50.1/52.3	**55.1/61.7**
Buddies	26.9/30.0	33.3/33.3	32.9/32.8	**34.9/34.0**	36.0/35.1	**41.5/41.1**

Table 3: *Overall accuracy/precision (%) for different baseline algorithms on the whole FAN-Large and interesting subsets. The experiments corresponding to last two columns (GBC (half) and GBC+CF (half)) is performed on the half of the dataset.*

each algorithm to be evaluated, except for the extended GBC (GBC+CF). For GBC+CF, we follow the procedure outlined in the previous section, i.e. we randomly divide the dataset into three subsets: training (40% of all the data), validation (10%) and test (50%) sets. The training and validation sets are used to train the contextual feature classifiers. After this training stage, the name-face assignment experiment is performed on the test set. For a fair comparison, and to determine whether contextual features help, we also rerun plain GBC on this smaller test set.

Results on the whole FAN-Large and on the interesting subsets are reported in table 3. Let us first consider the algorithms without caption contextual features. All algorithms using image content outperform the Random Assignment baseline. Constrained K-means achieves a lower accuracy compared to Constrained GMM and GBC, especially when the datasets are more difficult (*e.g.* the Hard and Buddies subsets), i.e. images have a higher average number of names per caption and "buddies" frequently appear together in the same image. It can also be observed that the difficulty of a dataset depends on the number of admissible assignments, which in turns depends on the average number of faces and names in an item. When the dataset is easy (*e.g.* Easy and Life.com), the gain in performance for the real approaches w.r.t the random baseline is high. Vice versa, on the hard datasets such as Hard and Buddies, the performance gains are more moderate. Another interesting observation which does not immediately appear from the table is that there is no strong correlation between the number of faces belonging to a name over the dataset and the accuracy of the assignments for that name (as long as there are more than just very few faces). The quality of the image is also an important factor: for example, on the Easy and Life subsets the average accuracy is significantly higher compared to the results obtained on the whole database. Among all algorithms, Constrained GMM is the most efficient, taking about 45 minutes to process the full database. GBC and Constrained K-means are more computational expensive (about 2.5 hours and 12 hours respectively).

After incorporating the caption contextual features into GBC, GBC+CF significantly outperforms GBC on the whole database as well as on all subsets (last two columns of table 3, tested on 50% of the data). Also, note how the performance of GBC on half the data is within 1% of the result on all the data. This suggests that GBC+CF delivers the best performance over *all* methods we compare. Two improvements are especially noteworthy. The data from the Life.com subset is collected from a single professional photography website, with high quality captions written by editors following predefined conventions. Therefore the contextual features provide GBC with very useful priors about whether a name appears in the image, resulting in a particularly strong improvement on the Life subset. Another interesting result comes from the Buddies subset. We found that in a large number of captions containing two Buddy names, only one of them appears in the image. The other name is mentioned for the completeness of the news story, but it often does not appear in the image. An example is given in fig. 6, where to the two names Barack Obama and Hillary Clinton are in the

"Buddies" list. In this particular case, our contextual classifier predicts the probabilities that Obama, Edward and Clinton appear in the image as 57.2%, 48.1% and 16.4%, respectively. Therefore, Clinton is the least likely to be assigned to a face, which helps GBC+CF to make the right assignment.

7 Conclusion

We presented a new large-scale database of images and captions, coined FAN-Large, for the automatic face naming task. It is designed to be representative of the challenges in learning automatically from realistic image and caption pairs mined freely from the Internet. Extensive experiments on the whole and various subsets of the database show the merits of our database. We also presented caption contextual features and shown how to incorporate them into a state-of-the-art face naming algorithm [12], thereby improving its performance. The proposed dataset could also be valuable in other studies such as verb and body pose association. We hope FAN-Large will become an important resource for research on image and captions in the computer vision community.

Acknowledgment This work was done while M. Özcan was an intern at the Idiap Research Institute. L. Jie was supported by PASCAL Pump Priming SS2-Rob Project, V. Ferrari was supported by a SNSF Professorship, and B. Caputo was supported by the SNSF project NINAPRO.

References

[1] http://opencv.willowgarage.com/wiki.

[2] http://opennlp.sourceforge.net.

[3] http://torch3vision.idiap.ch.

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

[5] T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In *CVPR*, 2009.

[6] N. Cristianini and J. S. Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[8] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.

[9] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy - automatic naming of characters in tv video. In *BMVC*, 2006.

[10] A. Farhadi, M. Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences for images. In *ECCV*, 2010.

[11] D. Forsyth, T. Berg, C. Alm, A. Farhadi, J. Hockenmaier, N. Loeff, and G. Wang. Words and pictures: Categories, modifiers, depiction and iconography. In *Object Categorization: Computer and Human Vision Perspectives, CUP*, 2009.

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

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

[14] A. Gupta and L. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008.

[15] S. J. Hwang and K. Grauman. Reading between the lines: Object localization using implicit cues from image tags. In *CVPR*, 2010.

[16] L. Jie and F. Orabona. Learning from candidate labeling sets. In *NIPS*, 2010.

[17] 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.

[18] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, 2007.

[19] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *PAMI*, 30(11):1958–1970, 2008.

[20] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. In *BMVC*, 2009.

[21] Y. Wang and G. Mori. A discriminative latent model of image region and object tag correspondence. In *NIPS*, 2010.