# Creating a Big Data Resource from the Faces of Wikipedia

**Md. Kamrul Hasan**
Génie informatique et génie logiciel
École Polytechnique de Montreal
Québec, Canada
md-kamrul.hasan@polymtl.ca

**Christopher J. Pal**
Génie informatique et génie logiciel
École Polytechnique de Montreal
Québec, Canada
christopher.pal@polymtl.ca

## Abstract

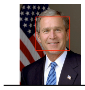
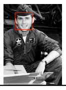
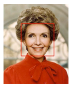
We present the *Faces of Wikipedia* data set in which we have used Wikipedia to create a large database of identities and faces. To automatically extract faces for over 50,000 identities we have developed a state of the art face extraction pipeline and a novel facial co-reference technique. Our approach is based on graphical models and uses the text of Wikipedia pages, face attributes and similarities, as well as clues from various other sources. Our method resolves the name-face association problem jointly for all detected faces on a Wikipedia page. We provide this dataset to the community for further research in various forms including: manually labeled faces, automatically labeled faces using our co-reference technique, raw and processed faces as well as text and meta data features for further evaluations of extraction and co-reference methods.

## 1 Introduction

Wikipedia is the largest and most diverse encyclopedia in human history and one of the top ten most popular sites on the Internet. About 15% of this encyclopedia covers human biographies. Wikipedia is constantly growing and being updated with new identities, textual content and facial images. For these and many other reasons, these biography pages provide an excellent source of raw data to explore data mining algorithms and to produce a big data resource for computer vision experiments. Vision data sets mined from the web have had substantial impact in recent years. For example the 80-million tiny images [1], the Caltech datasets [2], and ImageNet [3] are well known and widely used data sets. Other work, mining faces from news photos has lead to the well known Labeled Faces in the Wild (LFW) dataset and evaluation benchmark [4]. The original work that lead to the LFW also sought to automate the process of extracting faces automatically using both Natural Language Processing (NLP) and vision techniques [5]. Our work here is similar in spirit, but our mining task is different in a number of respects and we outline a few of them here. Firstly, the text captioning of Wikipedia images is not as standardized as the press photo captions that were used in [5]. Secondly, Wikipedia pages are structured documents with various other useful clues concerning the underlying content of images. Third, we wish to resolve all the faces, detected on all images from a Wikipedia biography page. As we shall see, we are able to exploit this aspect of the Wikipedia biography face mining problem to further increase extraction performance.

We introduce our face mining problem through an example. Consider the page of former president George W. Bush. Our mining goal is to classify all faces detected within the images of this page as either positive or negative examples of George W. Bush. The LFW evaluations focus on face verification, and results for face verification in the wild have improved dramatically over the past few years. Using such techniques, if we had a reference face of president Bush, one can imagine that during our mining task we could simply compare each test face with this known face, and decide whether they are same or not, i.e. use a pair-wise verification strategy for each image. In

Table 1: Wikipedia images with partial or full name match(in bold face), and noisy names (in Italic text)

| image | name : caption text | image | name : caption text |
|---|---|---|---|
|  | George W. Bush: 43rd President of the United States. |  | George W. Bush: Lt. **George W. Bush** while in the Texas Air National Guard. |
|  | Nancy Regan: Official White House photograph of **Nancy Reagan**, wife to then-President of the United States ***Ronald Reagan***. | | Omar Khadr: ***Rewakowski*** and ***Worth*** convalescing in hospital from their grenade injuries. |
| | Preity Zinta: **Zinta** as the teenage single mother ***Priya Bakshi*** in Kya Kehna (2000) which earned the actress her first nomination for Best Actress at Filmfare. | | George W. Bush: **Bush** thanks American military personnel, September 2007. |

our work here we are interested in extracting faces automatically without using any prior reference face information. In fact, this is one of the advantages of Wikipedia's biography page format. The simple existence of a biography page for a given person means faces on the page are likely to be the person of interest. Biography pages with a single face image contain a face of the person of interest 93% of the time based on our initial sampling and analysis. However, the situation quickly becomes more complex; for example, if an image caption says, "43rd President of the United States", as for the image in table 1, row 1 column 1, and we know who the 43rd US president is, we can infer that this face is referring to whom (George W. Bush). The second image in row 1 has a caption text as " Lt. *George W. Bush* while in the Texas Air National Guard". Clearly if we have a Named Entity Detector (NED) that can automatically detect the person name(s) in an image caption, it can give us important clue about who the person in the image is. Previous work has certainly used such information and we do here as well. The previous two examples are quite easy as they have an one to one name face and/or identity correspondence. Of course there is much more variability than this when we look at all the identities in the living people category of Wikipedia. Although our system has detected a face in the first image in row 2, there are two person names in the caption text. In the second image from row two, we have detected two faces and two names; however, none of them are true to our person of interest. The last row images are much more difficult - the left one has three faces and two person names, both names referring to the same person while the face detector fired 14 times in the right image. Using traditional NLP techniques we can resolve many of these ambiguities; however, the fact that we have multiple faces detected in multiple images allows us to combine NLP co-reference techniques with an approach to visual co-reference into one coherent model.

## 2 Proposed model

We now formalize the Wikipedia face mining problem definition as follows: consider processing the Wikipedia biography page of an identity, where we find $M$ images of at least a certain size. For each image, we run a face detector, and find $N_m$ faces (of at least a certain resolution). We define the faces as $\{\{x_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}$, where, $x_{mn}$ is the $n^{th}$ face from the $m^{th}$ image. The mining task is then to extract the true faces of our identity of interest, if any are found.

We model the problem with a graphical model with the structure shown in figure 1. The model contains the following set of random variables $\{\{\{X_{mn}, Y_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}, \{D_l\}_{l=1}^{L}, \{S_m\}_{m=1}^{M}, \}$, where, $X_{mn} = [X_1^{(mn)}, X_2^{(mn)}, \cdots X_K^{(mn)}]^T$ is the local feature vector for a face, $x_{mn}$, where $X_k^{(mn)}$ is the $k^{th}$ per face feature, described briefly below; $Y_{mn}$ is the binary output label $\{$1: true face, 0: false face$\}$; $S_m$ is a binary configuration constraint variable connecting output label variables, $\{Y_{mn}\}_{n=1}^{N_m}$, for all faces from image $m$. For a quantized similarity space, $D_l$ is the discrete pair-wise visual similarity variable, representing the bin index for a pair of faces, $x_{m_l'n_l'}$ and $x_{m_l''n_l''}$. The two faces forming a pair must be from two different images, and we have $L$ such pairs, $\{x_{m_l'n_l'}, x_{m_l''n_l''}\}_{l=1}^{L}$.

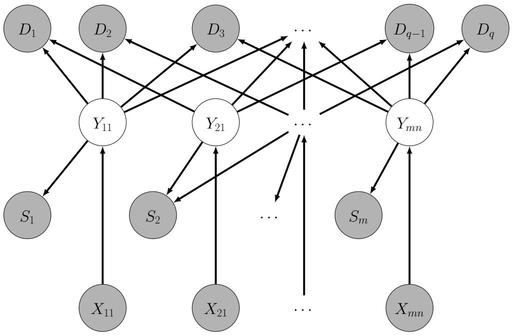


Figure 1: Our facial co-reference model.

In this setup, our facial identity resolution problem becomes an inference problem : the Maximum Probably Explanation (MPE), $\arg_Y \max p(Y|\{\{X_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}, \{D_l\}_{l=1}^{L}, \{S_m\}_{m=1}^{M})$ , where $Y = \{\{Y_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}$; Here,

$$
\begin{aligned}
&p(Y|\{\{X_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}, \{D_l\}_{l=1}^{L}, \{S_m\}_{m=1}^{M}) \propto p(Y, \{\{X_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}, \{D_l\}_{l=1}^{L}, \{S_m\}_{m=1}^{M}) \\
&\propto p(\{\{Y_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}, \{\{X_{mn}\}_{n=1}^{N_m}\}_{m=1}^{M}, \{D_l\}_{l=1}^{L}, \{S_m\}_{m=1}^{M}) \\
&\propto \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(Y_{mn}|X_{mn}) p(S_m|\{Y_{mn'}\}_{n'=1}^{N'_m}) \prod_{l=1}^{L} p(D_l|\{Y_{m'_l n'_l}, Y_{m''_l n''_l}\})
\end{aligned}
$$

In our joint model we use a discriminative maximum entropy classifier to model $p(Y_{mn}|X_{mn})$. For an input face pair, $\{x_{m'_l n'_l}\ , x_{m''_l n''_l}\}$, the corresponding output pair, $\{Y_{m'_l n'_l}, Y_{m''_l n''_l}\}$ has four possible discrete states: $\{\{1,1\}$ : both faces are true examples of our target identity, $\{1,0\}$ : first one is a true example, $\{0,1\}$ : second one is true, and, $\{0,0\}$ : neither is true$\}$. Here, we can model $\{1,0\}$ and $\{0,1\}$ configurations as a single never-same distribution, and therefore, we modeled the face-pair visual similarity distribution, $p(D_l|\{Y_{m'_l n'_l}, Y_{m''_l n''_l}\})$, as a conditional distribution, based on these three distribution definitions, $\{\{1,1\}$ : same (both are true), $\{\{0,1\},\{1,0\}\}$ : never same (one is true and the other is false), $\{0,0\}$ : rarely same (both are false)$\}$. We used a discrete distribution in the cosine space on the quantized cosine distances as the cosine based face verifiers [6] are among the leading performers in the Labeled Faces in the Wild (LFW) evaluations. To learn the proposed three class distributions we used the LFW view2 dataset - the same and the never-same classes are modeled through the 3000 positive and negative pairs, while the rarely same class is modeled through a weighted combination of positives and negatives with weights $0.25$ and $0.75$ respectively.

The idea of the binomial configuration constraint distribution, $p(S_m|\{Y_{mn}\}_{n=1}^{N_m})$, is to explicitly capture a specific constraint - it is less likely that two faces of the same individual appear on the same image. For faces, $\{x_{mn}\}_{n=1}^{N_m}$, that share the same parent image, $m$, the configuration constraint distribution is modeled as follows: when two or more faces in $\{x_{mn}\}_{n=1}^{N_m}$ are from the same identity, it is modeled through a Bernoulli distribution as,

$$
p(S_m|\{Y_{mn}\}_{n=1}^{N_m}) = \begin{cases} q & \text{when } S_m = 1 \\ 1-q & \text{otherwise} \end{cases}
$$

The distribution parameters, $q$, is learned from labeled instances via the Maximum Likelihood principle , i.e. $q = S_n/S_N$ , where, $S_n$ is the number of images with at least two faces from the same individual; and $S_N$ the total number of training images. When we have a single face from an image, i.e. $|N_m| = 1$, or no two faces in $\{x_{mn}\}_{n=1}^{N_m}$ are from the same identity, the configuration constraint distribution is modeled as always true, i.e.

Table 2: Prediction accuracy (%) & standard error for people with **2** faces (Acc $\pm S_E$). UT: unigrams (text), UTMI: unigrams (text + meta + image), UTMIB: unigrams (text + meta + image) + bigrams

| model | features | face count groups | | | | | | | Avg |
|---|---|---|---|---|---|---|---|---|---|
| | | 2 | 3 | 4 | 5 | 6 | 7 | $\geq 8$ | |
| | % true positives | 61 | 53 | 42 | 36 | 33 | 36 | 29 | 41 |
| MEM | UT | 69±4 | 65±4 | 65±3 | 60±3 | 64±4 | 63±3 | 65±2 | 64±3 |
| | UTMI | 71±3 | 75±3 | 71±3 | 66±3 | 72±4 | 68±4 | 69±2 | 70±3 |
| | UTMIB | 72±4 | 74±2 | 74±4 | 67±3 | 69±3 | 67±3 | 72±2 | 71±3 |
| our model | UT | 73±3 | 69±4 | 69±4 | 66±5 | 64±3 | 66±4 | 67±6 | 68±4 |
| | UTMI | 75±4 | 78±5 | 72±3 | 69±5 | 75±5 | 72±3 | 73±4 | 73±4 |
| | UTMIB | 78±5 | 80±4 | 77±4 | 70±4 | 71±3 | 73±2 | 74±6 | 75±4 |

$$p(S_m|\{Y_{mn}\}_{n=1}^{N_m}) = \begin{cases} 1 & \text{when } S_m = 1 \\ 0 & \text{otherwise} \end{cases}$$

The brute force search for the Most Probable Explanation (MPE) for $Y$ gets lengthy when the size of $Y$ is large. To deal with longer sequences, we have developed a chunk-based resolution protocol: for a sequence of size $\geq 8$, it is resolved through chunks with a chunk size of 7. Chunks are selected randomly and are provided with the currently most probable two faces as pivot assignments from earlier step(s). We initialize with the two most confident resolutions from the MEM classifiers.

## 3 Experiments and Discussion

As a part of this project, a spider was developed that traversed 522986 Wikipedia pages and downloaded 214869 images and the corresponding caption texts, if it found any. In addition to downloading images and caption texts, the spider was also asked to collect some meta information, for example, the location of an image in the Wikipedia page. As a next step, a face detector [7] was used to extract faces. For each detection, the faces were cut out from images with an additional 1/3 background. Thus, we had a total of 90,453 faces, out of which 51300 were from people with only one face, and 12991 people had at least two faces. For person name detection in the caption text, we used the Stanford Named Entity Detector (NED)[8]. We grouped and sampled the Wikipedia people based on the number of faces for labeling. From single face people we randomly selected 250 identities, while for groups 2-7, we sampled 100 identities each, and for group 8 (the group with $\geq$ 8 faces), we labeled all their faces. Among the 9743 sampled faces we had 3466 true, 5324 false, and 953 noisy faces. Here, by noisy face we mean either no face was detected in the image or the image was not a photograph of a face (ex. it was a cartoon, drawing, sketch, or sculpture etc.)

A set of unigrams and bigrams define the feature set, $\{\{X_k^{(mn)}\}\}_{k=1}^{K}$, for a face, $x_{mn}$. While unigrams are independent local features, bigrams are the logical anding of any two unigrams. In particular, we carefully defined and tested a set of unigram features and selected a subset from those to generate the bigrams. The unigrams are from three different sources: (i) Text features: from caption text of an image and the image file name; for example, the person name, whether appeared in the caption text or in the image file name; certain linguistic token(s), if detected in those texts. (ii) Image and meta features: for example, the location of the image (info box or from other locations), the number of faces, detected in the image. (iii) Face features: for example, the portion of image area covered by the face, relative size of the face, compared to the others from the same image. We have used Local Binary Pattern (LBP) [9] features as the definition of a face. Before extracting LBP features, the faces went through a face alignment and the patch selection pipeline as approached in [10]. The LBP feature vectors of size 7080 were projected through two projections, each of which reduce their dimensionality: first, we use a PCA projection down to 500 dimensions, this is followed by a projection using a variation of the cosine distance metric learning approach discussed in [11]. The final feature vectors were of size 200. Among the labeled 250 single face instances 7% were noisy (not photographs of faces), but among the non-noisy faces 98.5% were true positives. Thus, a true single face detection in a Wikipedia human biography site can be inferred as a true example of the person of interest with high precision. Table 2 summarizes the results and compares our model with a MEM for various different feature types. For each group, a randomly chosen 70% of the labeled instances plus its 1 neighbor group(s) (for example, group 3 and 4 are the 1 neighbors of group 2) were used for training while the rest 30% were used for testing. Our model had a 4% gain over a MEM and for both the models, the UTMI features improved the accuracy by (5-6)% over the UT features. The addition of bigrams gave us a further boost by 2%. We see that the combination of our joint inference technique and more sophisticated feature engineering are able to boost our extraction accuracy substantially.

Analyzing the labeled data that we have obtained so far, we might expect approximately 56000 true faces, with roughly 47,700 from single face identities and 8300 from people with at least two faces in their Wikipedia page. The prominent LFW dataset provides us with 13233 faces for 5749 identifies. We have the ability to produce a data set that is approximately 4 times larger than the LFW in terms of faces and 9 times larger data set in terms of identity count. To the best of our knowledge this would represent one of the largest publicly available face data sets. We intend to incrementally release as much human labeled data as possible, using our facial co-reference techniques to focus labelling energy on higher uncertainly cases. Additionally, because of the dynamics of Wikipedia database we would be able to boost the database on a timely basis with minimum cost and effort.

## Acknowledgements

We thank Erik Learned-Miller and Yoshua Bengio for discussion on various aspects of this work.

## References

[1] Antonio Torralba, Robert Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970, 2008.

[2] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[4] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[5] Tamara L. Berg, Er C. Berg, Jaety Edwards, Michael Maire, Ryan White, Yee whye Teh, Erik Learned-miller, and D. A. Forsyth. Names and faces in the news. In *In Proc. CVPR*, pages 848–854. IEEE Computer Society, 2004.

[6] Hieu V. Nguyen and Li Bai. Cosine similarity metric learning for face verification. In *ACCV (2)*, pages 709–720, 2010.

[7] Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.

[8] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[9] Pietikinen M Ojala T and Menp T. A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification. *Proc. of CAPR, Rio de Janeiro, Brazil.*, (2013):397–406, 2001.

[10] M. K. Hasan and C. Pal. Improving Alignment of Faces for Recognition. In *IEEE Symposium on Robotic and Sensors Environments (ROSE)*, pages pp. 249–254, 2011.

[11] A. Anonymous. Improving face verification in the wild with poses, and scaling up recognition with discriminative metric data structures (in review). *International Journal of Multimedia Information Retrieval*, 2012.