Names and Faces

Tamara L. Berg, Alexander C. Berg, Jaety Edwards, Michael Maire,
Ryan White, Yee-Whye Teh, Erik Learned-Miller, D.A. Forsyth

University of California, Berkeley
Department of Computer Science
Berkeley, CA 94720

Abstract

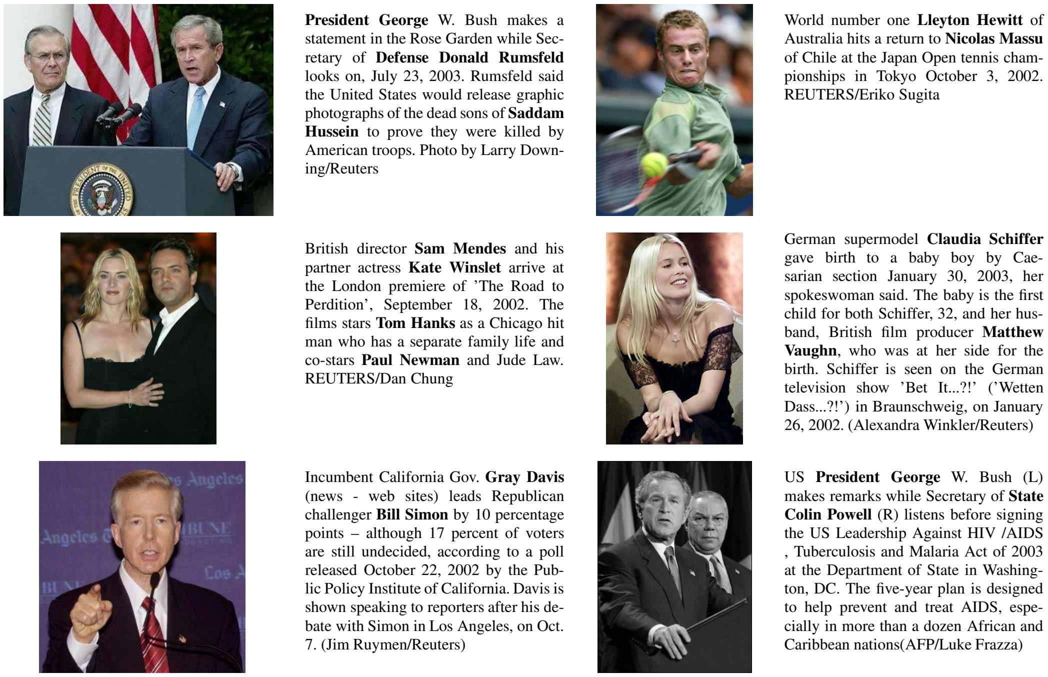
We show that a large and realistic face dataset can be built from news photographs and their associated captions. Our automatically constructed face dataset consists of 30,281 face images, obtained by applying a face finder to approximately half a million captioned news images and labeled using image information from the photographs and word information extracted from the corresponding caption. This dataset is more realistic than usual face recognition datasets, because it contains faces captured “in the wild” in a variety of configurations with respect to the camera, taking a variety of expressions, and under illumination of widely varying color. Faces are extracted from the images and names with context are extracted from the associated caption. Our system uses a clustering procedure to find the correspondence between faces and associated names in news picture-caption pairs.

The context in which a name appears in a caption provides powerful cues as to whether it is depicted in the associated image. By incorporating simple natural language techniques, we are able to improve our name assignment significantly. We use two models of word context, a Naive Bayes model and a Maximum Entropy model. Once our procedure is complete, we have an accurately labeled set of faces, an appearance model for each individual depicted, and a natural language model that can produce accurate results on captions in isolation.

Key words: Names; Faces; News; Words; Pictures;

1 Introduction

There are many datasets of images with associated words. Examples include: collections of museum material; the Corel collection of images; any video with sound or closed captioning; images collected from the web with their enclosing web pages; or captioned news images.



President George W. Bush makes a statement in the Rose Garden while Secretary of **Defense Donald Rumsfeld** looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of **Saddam Hussein** to prove they were killed by American troops. Photo by Larry Downing/Reuters

World number one **Lleyton Hewitt** of Australia hits a return to **Nicolas Massu** of Chile at the Japan Open tennis championships in Tokyo October 3, 2002. REUTERS/Eriko Sugita

British director **Sam Mendes** and his partner actress **Kate Winslet** arrive at the London premiere of 'The Road to Perdition', September 18, 2002. The films stars **Tom Hanks** as a Chicago hit man who has a separate family life and co-stars **Paul Newman** and Jude Law. REUTERS/Dan Chung

German supermodel **Claudia Schiffer** gave birth to a baby boy by Caesarian section January 30, 2003, her spokeswoman said. The baby is the first child for both Schiffer, 32, and her husband, British film producer **Matthew Vaughn**, who was at her side for the birth. Schiffer is seen on the German television show 'Bet It...?!' ('Wetten Dass...?!') in Braunschweig, on January 26, 2002. (Alexandra Winkler/Reuters)

Incumbent California Gov. **Gray Davis** (news - web sites) leads Republican challenger **Bill Simon** by 10 percentage points – although 17 percent of voters are still undecided, according to a poll released October 22, 2002 by the Public Policy Institute of California. Davis is shown speaking to reporters after his debate with Simon in Los Angeles, on Oct. 7. (Jim Ruymen/Reuters)

US **President George** W. Bush (L) makes remarks while Secretary of **State Colin Powell** (R) listens before signing the US Leadership Against HIV /AIDS , Tuberculosis and Malaria Act of 2003 at the Department of State in Washington, DC. The five-year plan is designed to help prevent and treat AIDS, especially in more than a dozen African and Caribbean nations(AFP/Luke Frazza)

Fig. 1. *Some typical news photographs with associated captions from our dataset. Notice that multiple faces may appear in the pictures and multiple names may occur in the associated captions. Our task is to detect faces in these pictures, detect names in the associated captions and then correctly label the faces with names (or "NULL" if the correct name does not appear in the caption or the named entity recognizer does not detect the correct name). The output of our system on these images appears in figure 5.*

It is a remarkable fact that pictures and their associated annotations are complimentary. This observation has been used to browse museum collections (4), and organize large image collections. In particular, several models have been used to organize the Corel Dataset of images with associated keywords. Barnard *et al* (3) used a multi-modal extension to mixture of latent Dirichlet allocation to predict words associated with whole images as well as words corresponding to particular image regions in an auto-annotation task. Li and Wang (22) used 2-dimensional multi-resolution hidden markov models on categorized images to train models representing a set of concepts. They then used these concepts for automatic linguistic indexing of pictures. Lavrenko *et al* (20) used continuous space relevance models to predict the probability of generating a word given image regions for automatic image annotation and retrieval. Words and pictures have also been combined to perform other tasks such as image segmentation (11), and object recognition (13).

Previous research has produced quite good results by exploiting the complimentary nature of words and pictures, but has relied on relatively simple image and word representations. All of the previously-mentioned papers have represented images as regions found through various forms of low-level segmentation. In this work we exploit the past decade of computer vision research building specialized detectors for certain classes of objects and focus on faces in images. Faces are the best example of a domain where object detection has been successful and very good face detectors are available e.g. (23; 32; 37); we use the detector of (23).

The work most similar to ours is that of Fitzgibbon and Zisserman (16). They automatically discover cast listings in video using affine-invariant clustering methods on detected faces and are robust to changes in lighting, viewpoint and pose. More recently Arandjelovic and Zisserman (1) have extended this work to suppress effects of background surrounding the face, refine registration and allow for partial occlusion and expression change. Their work has many of the same image based challenges that we do, but operates in a slightly different regime than ours; they are working with faces in video whereas we have still frame photographs and captions.

Concentrating on objects in images, in particular faces, provides the motivation to similarly emphasize certain parts of the associated text – named entities. Research in natural language processing has produced useful named entity recognizers, which can identify specific substrings, proper names, in captions that may refer to faces in the associated image.

Our basic task is to find a correspondence between the names and faces. A correspondence is in fact a labeling of the faces. The set of correspondences allows us to build a model for each individual's appearance (from their set of labeled faces). In addition the correspondences provide training data for a natural language model that recognizes what context around a name indicates it will be pictured, and possibly how it will be pictured. Using these learned appearance and language models the estimated correspondence can be improved. In this paper, solving the correspond problem and fitting the associated appearance and natural language models are combined in an iterative alternating optimization framework.

1.1 Faces

Face recognition is a well studied problem. Early recognition techniques used nearest neighbor classifiers based on pixel values. The nearest neighbor search was made more efficient and possibly robust using dimensionality reduction called Eigenfaces (31; 35). Later, linear discriminant methods were proposed (5) that utilized class information to produce an improved distance metric and better recognition results. More recently, it has been found that models based on 3D structure, lighting, and surface appearance (10; 24) or appearance based methods that explicitly model pose (19) give better recognition accuracy, but can be somewhat hard to fit for arbitrary faces. Some reviews of face recognition methods appear in (18; 40; 24). Our goal is more restricted than general face recognition, in that we need only distinguish between a small number of names in the corresponding caption. This is a significant simplification of the face recognition task. As a result, we can use a fairly simple face representation as discussed in section 3.3.

Doctor Nikola shows a fork that was removed from an Israeli woman who swallowed it while trying to catch a bug that flew in to her mouth, in Poriah Hospital northern Israel July 10, 2003. Doctors performed emergency surgery and removed the fork. (Reuters)

President George W. Bush waves as he leaves the White House for a day trip to North Carolina, July 25, 2002. A White House spokesman said that Bush would be compelled to veto Senate legislation creating a new department of homeland security unless changes are made. (Kevin Lamarque/Reuters)

Fig. 2. *In the news dataset a few individuals, like President Bush (***right***), appear frequently in the news so we have many pictures of them. Whereas most people, like Dr. Nikola (***left***) appear only a few times or in only one picture. This distribution reflects what we would expect from real applications. For example, in airport security cameras, a few people, (e.g. airline staff) might be seen often, but the majority of people would appear infrequently. Studying how recognition systems perform under these circumstances and providing datasets with these features is necessary for producing reliable face recognition systems.*

As can be seen in the literature, faces are difficult to recognize. Although face recognition is well studied, the disparity between results reported in research papers and in real world field tests of recognition systems is quite large (29). It has been shown (25) that the performance of a face recognition system on a dataset can largely be predicted by the performance of a baseline algorithm, such as principal component analysis, on the same dataset. Since recognition systems work well on current face datasets, but poorly in practice, this suggests that the datasets currently used are not representative of real world settings. Because current datasets were captured in the lab, they may lack important phenomena that occur in real face images. To solve face recognition, systems will have to deal well with a dataset that is more realistic, with wide variations in color, lighting, expression, hairstyle and elapsed time.

One consequence of our work is a labeled dataset captured “in the wild” consisting of faces from news photographs. This dataset displays the phenomena found in real world face recognition tasks and is derived from a large collection of news photographs with associated captions collected from the world wide web at a rate of hundreds to over a thousand per day. While the images are captioned, the identity of individual faces is not given. Many images contain multiple faces, and the associated captions contain many names. In this paper we show good solutions to this correspondence problem, resulting in a face dataset that is measurably more challenging than current face recognition datasets (section 6).

The process for building our face dataset consists of: detecting names using the open source named entity recognizer of (12), detecting and representing faces (section 3), and then associating those names with faces. Initially we use a basic clustering method to assign names to faces (first described in our CVPR 2004 paper (7)). This produces quite a good clustering. However, it ignores important language information that can be used to produce even better results. For example, the named entity recognizer occasionally identifies names that do not correspond to actual people (e.g. "U.S. Open"). In section 5 – and our previous work (8) – we show that by incorporating simple natural language techniques we can determine the probability of a name being pictured in the corresponding picture and use this information to improve our results significantly. An attractive byproduct of our system is a natural language module which can be used to analyze text in isolation. In this work, we implement two models of context a Naive Bayes model and a Maximum Entopy model and compare their performance.

2 News Dataset

We have collected a dataset consisting of approximately half a million news pictures and captions from Yahoo News over a period of roughly two years. Using Mikolajczyk's face detector (23), we extract faces from these images; using Cunningham *et al*'s open source named entity recognizer (12), we detect proper names in each of the associated captions. This gives us a set of faces and names associated with each picture. Our task is to assign one of these names or null (unnamed) to each detected face.

Our dataset differs from typical face recognition datasets in a number of important ways:

Pose, expression and illumination vary widely. The face detector tends not to detect lateral views of faces, but we often encounter the same face illuminated with markedly different colored light and in a very broad range of expressions. Spectacles and mustaches are common (Figure 6). There are wigs, images of faces on posters, differences in resolution and identikit pictures (e.g. Figure 6). Quite often there are multiple copies of the same picture (this is due to the way news pictures are prepared, rather than a collecting problem) or multiple pictures of the same individual in similar configurations. Finally, many individuals are tracked across time which has been shown to hamper face recognition substantially (18).

Name frequencies have the long tails that occur in natural language problems. We expect that face images follow roughly the same distribution. We have hundreds to thousands of images of a few individuals (e.g. *President Bush*), and a large number

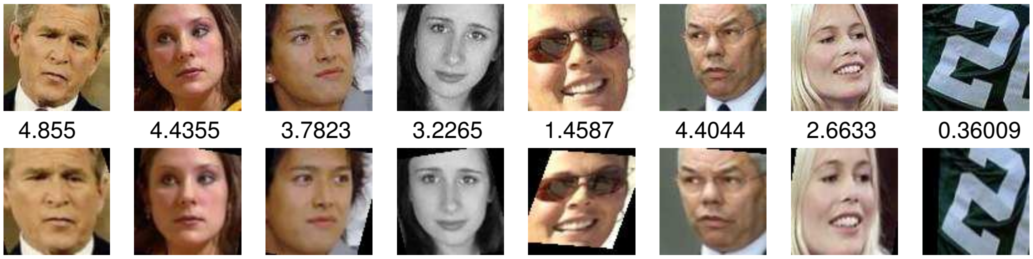


Fig. 3. *The face detector can detect faces in a range of orientations, as the* **top row** *shows. Before clustering the face images we rectify them to a canonical pose* **bottom row**. *The faces are rectified using a set of SVM's trained to detect feature points on each face. Using gradient descent on SVM outputs, the best affine transformation is found to map detected feature points to canonical locations. Final rectification scores for each of these faces are shown* **center** *(where larger scores indicate better performance). This means that incorrect detections, like the rightmost image can be discarded because of their poor rectification scores.*

of individuals who appear only a few times or in only one picture (e.g. Figure 2). One expects real applications to have this property. For example, in airport security cameras a few people, security guards, or airline staff might be seen often, but the majority of people would appear infrequently. Studying how recognition systems perform under these circumstances is important.

The sheer **volume** of available data is extraordinary. We have sharply reduced the number of face images we deal with by using a face detector that is biased to frontal faces and by requiring that faces be large and rectify properly. Even so, we have a dataset that is comparable to, or larger than, the biggest available lab sets and is much richer in content. Computing kernel PCA and linear discriminants for a set this size requires special techniques (section 3.3.1).

3 Finding and Representing Faces

For each news picture we:

(1) Detect faces in the images (Section 3.1). We confine our activities to large, reliably detected faces, of which 44,773 are found.
(2) Rectify those faces to a canonical pose (Section 3.2). After throwing out poorly rectified faces, this further reduces our dataset to 34,623 faces.
(3) Transform the face into a representation suitable for the assignment task (Section 3.3).
(4) From these 34,623 faces we confine ourselves to faces with proper names detected in their corresponding captions, leaving 30,281 faces, the final set we run our assignment procedure on and the number of faces in the dataset that we produce.

3.1 Face detection

For face detection, we use Mikolajczyk's implementation (23) of the face detector described by Schneidermand and Kanade (32). To build this face detector, a training set of face and non-face images is used to determine the probability of a new image being a face. Each image in the training set is decomposed into a set of wavelet coefficients which are histogrammed so that each bin corresponds to a distinct set of coefficients. The probability of a new image being a face is the number of face images assigned compared to the number of non-face images assigned to its bin. This provides 44,773 large well detected face images (size 86x86 pixels or larger with sufficient face detection scores and resized to 86x86 pixels).

3.2 Rectification

Before comparing images, we use a novel method to automatically rectify all faces to a canonical pose. Rectification aligns images so that comparisons between faces are applied to corresponding parts of the face. We train five support vector machines as feature detectors for several features on the face (corners of the left and right eyes, corners of the mouth, and the tip of the nose) using a training set consisting of 150 hand clicked faces. We use the geometric blur of (6) applied to gray-scale patches as the features for our SVM. Using geometric blur features instead of raw image patches greatly increases rectification accuracy and was a necessary step to making our rectification system effective.

The geometric blur descriptor first produces sparse channels from the grey scale image, in this case, half-wave rectified oriented edge filter responses at three orientations yielding six channels. Each channel is blurred by a spatially varying Gaussian with a standard deviation proportional to the distance to the feature center. The descriptors are then sub-sampled and normalized. Initially image patches were used as input to the feature detectors, but replacing patches with the geometric blurred version of the patches produced quite significant gains in rectification accuracy.

A new face is rectified by computing each SVM output over the entire image with a weak prior on location for each feature. This produces a set of 5 feature maps where the value of the map at any pixel is the SVM output for that feature. Using the least squares solution between the maximal outputs of each SVM feature detector and the canonical feature locations, we compute an initial estimate for the affine transformation between the face and the canonical pose. This solution is further refined using gradient descent on the SVM feature maps to find the overall best affine transformation mapping detected points to canonical feature locations. Each image is rectified to a common pose using the computed affine transformations and assigned a score based on the sum of its feature detector responses (larger scores imply better rectification). Notice that errors in face detection (Figure 3) can be removed by thresholding on rectification score (center number – larger numbers indicate a better score).

We filter our dataset by removing images with poor rectification scores, leaving 34,623 face images. Each face is automatically cropped to a region surrounding the eyes, nose and mouth to eliminate effects of background on recognition. The RGB pixel values from each cropped face are concatenated into a vector and used from here on.

3.3 Face Representation

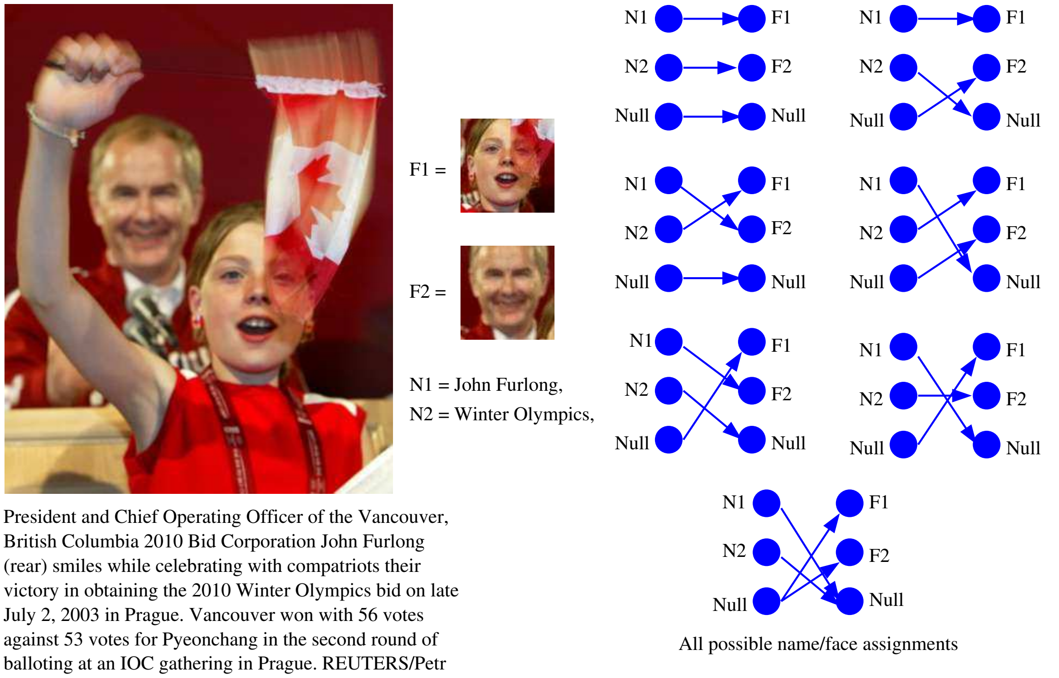
We model appearance using a mixture model with one mixture element per name in our lexicon, $P(face|name)$. Optimally, the face representation should be in a feature space where comparisons are helpful. To achieve a good feature space we first rectify the cropped face regions, then compute a basis using kernel principal components analysis (kPCA) for dimensionality reduction followed by linear discriminant analysis (LDA). LDA has been shown to work well for face discrimination (41; 5; 18) because it uses class information to find a set of discriminants that separate data points from different classes sufficiently. We model the distributions $P(face|name)$ using gaussians with fixed covariance in this feature space.

3.3.1 kPCA and the Nyström Approximation

Kernel Principal Components Analysis: Kernel Principal Components Analysis (kPCA) (33) uses a kernel function to efficiently compute a principal component basis in a high-dimensional feature space, related to the input space by some nonlinear map. Kernel PCA has been shown to perform better than PCA at face recognition (39). Kernel PCA is performed as follows:

- Compute a kernel matrix, K, where K_{ij} is the value of the kernel function comparing $image_i$ and $image_j$ (we use a Gaussian kernel with independent diagonal variances).
- Center the kernel matrix in feature space by subtracting off average row, average column and adding on average element values.
- Compute an eigendecomposition of K, and project onto the normalized eigenvectors of K.

Our dataset is too large to do kPCA directly as the kernel matrix K will be size NxN, where N is the the number of images in the dataset, and involve approximately 10^9 image comparisons. Therefore, we instead use an approximation to calculate the eigenvectors of K. Incomplete Cholesky Decomposition (ICD) can be used to calculate an approximation to K with a bound on the approximation error (2), but involves accessing all N images for each column computation (where N is



President and Chief Operating Officer of the Vancouver, British Columbia 2010 Bid Corporation John Furlong (rear) smiles while celebrating with compatriots their victory in obtaining the 2010 Winter Olympics bid on late July 2, 2003 in Prague. Vancouver won with 56 votes against 53 votes for Pyeonchang in the second round of balloting at an IOC gathering in Prague. REUTERS/Petr Josek

Fig. 4. *To assign faces to names, we evaluate all possible assignments of faces to names and choose either the maximum likelihood assignment or form an expected assignment. Here we show a typical data item (***left***), with its detected faces and names (***center***). The set of possible correspondences for this data item are shown at* **right***. This set is constrained by the fact that each face can have at most one name assigned to it and each name can have at most one face assigned, but any face or name can be assigned to Null. Our named entity recognizer occasionally detects phrases like "Winter Olympics" which do not correspond to actual people. These names are assigned low probability under our language model, making their assignment unlikely. EM iterates between computing the expected value of the set of possible face-name correspondences and updating the face clusters and language model. Unusually, we can afford to compute all possible face-name correspondences since the number of cases is small. For this item, we correctly choose the best matching "F1 to Null", and "F2 to N1".*

the number of images in the dataset, currently 34,623). This makes computation relatively lengthy.

The Nyström approximation method (cf (38; 17)) gives a similar result, but allows the images to be accessed only once in a single batch rather than once for each column computation, making it much faster to compute for large matrices than ICD.

The Nyström method computes two exact subsets of K, A and B, and uses these to approximate the rest of K. Using this approximation of K, the eigenvectors can be approximated efficiently.

First the N×N kernel matrix, K, is partitioned as

$$K = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

with $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{(N-n) \times n}$ and $C \in \mathbb{R}^{(N-n) \times (N-n)}$. Here, A is a subset of the images, (in our case 1000 randomly selected images) compared to themselves, B is the comparison of each of the images of A, to the rest of the images in our dataset, and C is approximated by the Nyström method. Nyström gives an approximation for C as, $\hat{C} = B^T A^{-1} B$. This gives an approximation to K, $\hat{K} = \begin{bmatrix} A & B \\ B^T & \hat{C} \end{bmatrix}$

Then we form $\tilde{K}$, the centered version of our approximation $\hat{K}$, by calculating approximate average row, average column sums (these are equal since K is symmetric), and average element values. We can approximate the average row (or column) sum as:

$$\hat{K} 1_N = \begin{bmatrix} A 1_n + B 1_{N-n} \\ B^T 1_n + B^T A^{-1} B 1_{N-n} \end{bmatrix}$$

We center as usual,

$$\tilde{K} = \hat{K} - \frac{1}{N} 1_N \hat{K} - \frac{1}{N} \hat{K} 1_N + \frac{1}{N^2} 1_N \hat{K} 1_N.$$

We solve for the orthogonalized approximate eigenvectors as follows. First, we replace A and B by their centered versions. Let $A^{\frac{1}{2}}$ be the square root of A, and $S = A + A^{-\frac{1}{2}} B B^T A^{-\frac{1}{2}}$. Diagonalize S as $S = U_s \Lambda_s U_s^T$. Then $\tilde{K}$ is diagonalized by:

$$V = \begin{bmatrix} A \\ B^T \end{bmatrix} A^{-\frac{1}{2}} U_s \Lambda_s^{-\frac{1}{2}}$$

Then we have $\tilde{K} = V \Lambda_s V^T$ and $V^T V = I$. Given this decomposition of $\tilde{K}$ we proceed as usual for kPCA, by normalizing the eigenvectors Λ_s and projecting $\tilde{K}$ onto the normalized eigenvectors. This gives a dimensionality reduction of our images that makes the discrimination task easier.

Nyström does not give the same error bound on its approximation to K as ICD. However, we expect the number of large eigenvalues of our matrix to be small as a result of the smoothing properties of kernel functions. This implies that the effective column rank of kernel matrices should be low. Therefore, we should be able to observe all of the column rank with a small subset of the columns and the approximation error should be small. In our matrix we observed that the eigenvalues of A do tend to drop off quickly. Because there is nothing special about this subset of faces (they were chosen at random from our set), the effective rank of the whole matrix should be small, and the Nyström method should provide a good approximation.

3.3.2 LDA

We use the usual definition of LDA (41; 5; 18) where each image belongs to a set of m classes $(C_1, ..., C_m)$, there are N_i samples in class i and μ_i is the mean of class i. Then the within and between class scatter matrices are defined as:

$$\begin{aligned} W &= \sum_{i=1}^{m} \sum_{x_j \in C_i} (x_j - \mu_i)(x_j - \mu_i)^T \\ B &= \sum_{i=1}^{m} N_i(\mu_i - \mu)(\mu_i - \mu)^T \end{aligned}$$

LDA computes the projection α that maximizes the ratio:

$$\alpha_{opt} = argmax_{\alpha} \frac{|\alpha^T B \alpha|}{|\alpha^T W \alpha|}$$

by solving the generalized eigenvalue problem:

$$B\alpha = \lambda W \alpha$$

In order to compute LDA, class labels must be known. However, we do not have a training set or any labeled data since ours is an unsupervised recognition task. To bypass this restriction, we use as proxy to labeled training data, the images from our dataset that contain only one detected face and that have only one detected name in the associated caption. This gives us a slightly noisy training set on which to compute the LDA vectors, but which is accurate enough to improve performance significantly on top of the space found by kPCA.

4 Name Assignment by Simple Clustering

Our model for generating a news item with F faces and N names, consists of first generating N names with contexts. For each of these names, a binary variable $pictured$ is generated. For every name with $pictured = 1$, a face is generated. Each of the remaining, $F - \sum pictured$ unnamed faces is generated according to a distribution $P(face)$.

A natural way of thinking about name assignment is as a hidden variable problem where the hidden variables are the correct name-face correspondences for each picture. This suggests using an expectation maximization (EM) procedure. EM iterates between computing an expectation over face-name correspondences (given a face clustering) and updating the face clusters. Unusually, it is affordable to enumerate and consider all of the face-name correspondences since the number of cases is small.

4.1 Name Assignment

For each picture, we calculate the likelihood of each possible assignment. Each name can be assigned to at most one face, each face can be assigned to at most one name and null can be assigned to any name or face. An example of the extracted names, faces and all possible assignments can be seen in figure 4.

We write P(N) as the probability of generating N names, and P(F) as the probability of generating F faces. For an assignment a_j (consisting of a correspondence of names to faces), letting α index into the names that are pictured, $\sigma(\alpha)$ index into the faces assigned to the pictured names, and γ index into the faces without assigned names, the likelihood of picture x_i under assignment a_j, of names to faces is:

$$L_{x_i,a_j} = P(N)P(F) * \prod_{\alpha} P(f_{\sigma(\alpha)}|n_\alpha) \prod_{\gamma} P(f_\gamma)$$

The terms $P(N)P(F)$ are independent of the assignment so can be dropped when calculating the probability of an assignment. We focus on the remaining terms to calculate assignment likelihoods.

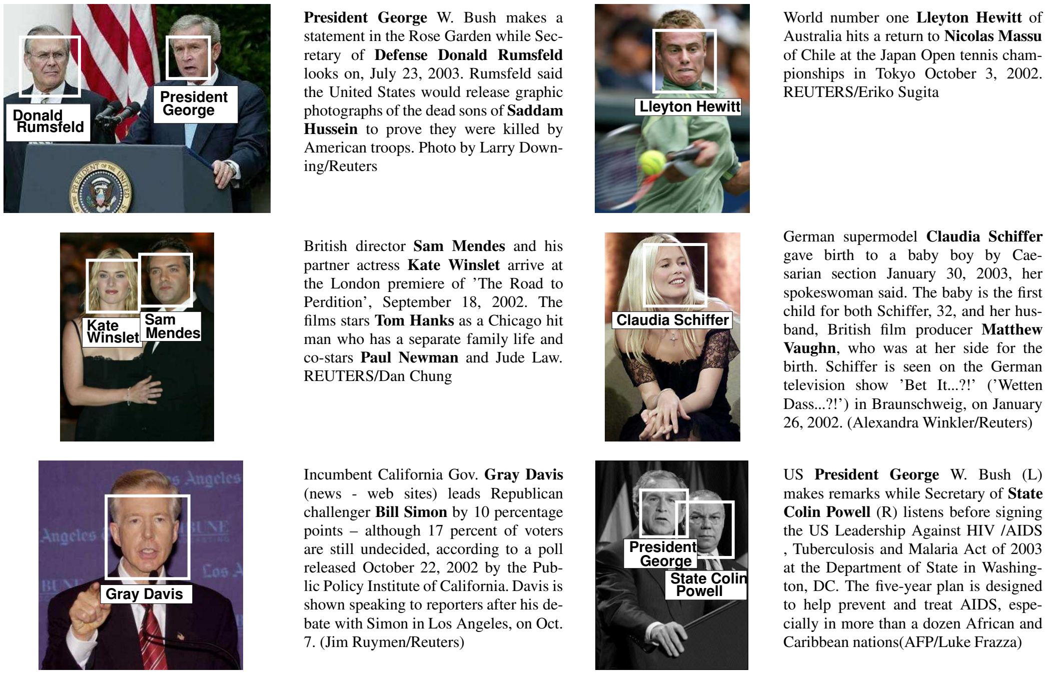
The complete data log likelihood is:

$$\sum_{i \epsilon pics} \left[\sum_{j \epsilon C_i} (\delta_{ij} log(L_{x_i,a_j}) \right]$$

Where C_i is the set of possible assignments for image i, and δ_{ij} is an indicator variable telling which of the available correspondences occurred in this data item. The δ_{ij} are missing data whose expectations are computed in the E step.

This gives a straightforward EM procedure:

- E – update the δ_{ij} according to the normalized probability of picture i with assignment j.
- M – maximize the parameters $P(face|name)$ using soft counts.



President George W. Bush makes a statement in the Rose Garden while Secretary of **Defense Donald Rumsfeld** looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of **Saddam Hussein** to prove they were killed by American troops. Photo by Larry Downing/Reuters

World number one **Lleyton Hewitt** of Australia hits a return to **Nicolas Massu** of Chile at the Japan Open tennis championships in Tokyo October 3, 2002. REUTERS/Eriko Sugita

British director **Sam Mendes** and his partner actress **Kate Winslet** arrive at the London premiere of 'The Road to Perdition', September 18, 2002. The films stars **Tom Hanks** as a Chicago hit man who has a separate family life and co-stars **Paul Newman** and Jude Law. REUTERS/Dan Chung

German supermodel **Claudia Schiffer** gave birth to a baby boy by Caesarian section January 30, 2003, her spokeswoman said. The baby is the first child for both Schiffer, 32, and her husband, British film producer **Matthew Vaughn**, who was at her side for the birth. Schiffer is seen on the German television show 'Bet It...?!' ('Wetten Dass...?!') in Braunschweig, on January 26, 2002. (Alexandra Winkler/Reuters)

Incumbent California Gov. **Gray Davis** (news - web sites) leads Republican challenger **Bill Simon** by 10 percentage points – although 17 percent of voters are still undecided, according to a poll released October 22, 2002 by the Public Policy Institute of California. Davis is shown speaking to reporters after his debate with Simon in Los Angeles, on Oct. 7. (Jim Ruymen/Reuters)

US **President George** W. Bush (L) makes remarks while Secretary of **State Colin Powell** (R) listens before signing the US Leadership Against HIV /AIDS , Tuberculosis and Malaria Act of 2003 at the Department of State in Washington, DC. The five-year plan is designed to help prevent and treat AIDS, especially in more than a dozen African and Caribbean nations(AFP/Luke Frazza)

Fig. 5. *Given an input image and an associated caption (images above and captions to the right of each image), our system automatically detects faces (white boxes) in the image and possible name strings (bold). We use a clustering procedure to build models of appearance for each name and then automatically label each of the detected faces with a name if one exists. These automatic labels are shown in boxes below the faces. Multiple faces may be detected and multiple names may be extracted, meaning we must determine who is who (e.g., the picture of* Claudia Schiffer*).*

4.2 Best Correspondence vs. Expected Correspondence

From a statistical point of view, given our hidden variable problem of determining name-face pairings, EM seems like the most favorable solution. However, it is known that for a variety of vision problems where one might reasonably expect EM to be a natural algorithm, searching over missing variables performs significantly better. The best known example occurs when one wishes to estimate the fundamental matrix relating two views of a scene. Here, missing variables identify which pairs of points correspond. The best known methods for solving this problem involve a form of randomized search over missing variables (RANSAC, first described in (15), and applied to this problem in (34); or MLESAC, a recent variant (36)) and significantly outperform EM on this problem. These methods choose the assignment that maximizes the complete data log-likelihood, rather than taking an expectation over missing assignments.

The Maximal Assignment process is quite similar to the EM process except instead of calculating the expected value of each assignment, only the maximum likelihood assignment is given a nonzero probability of 1.

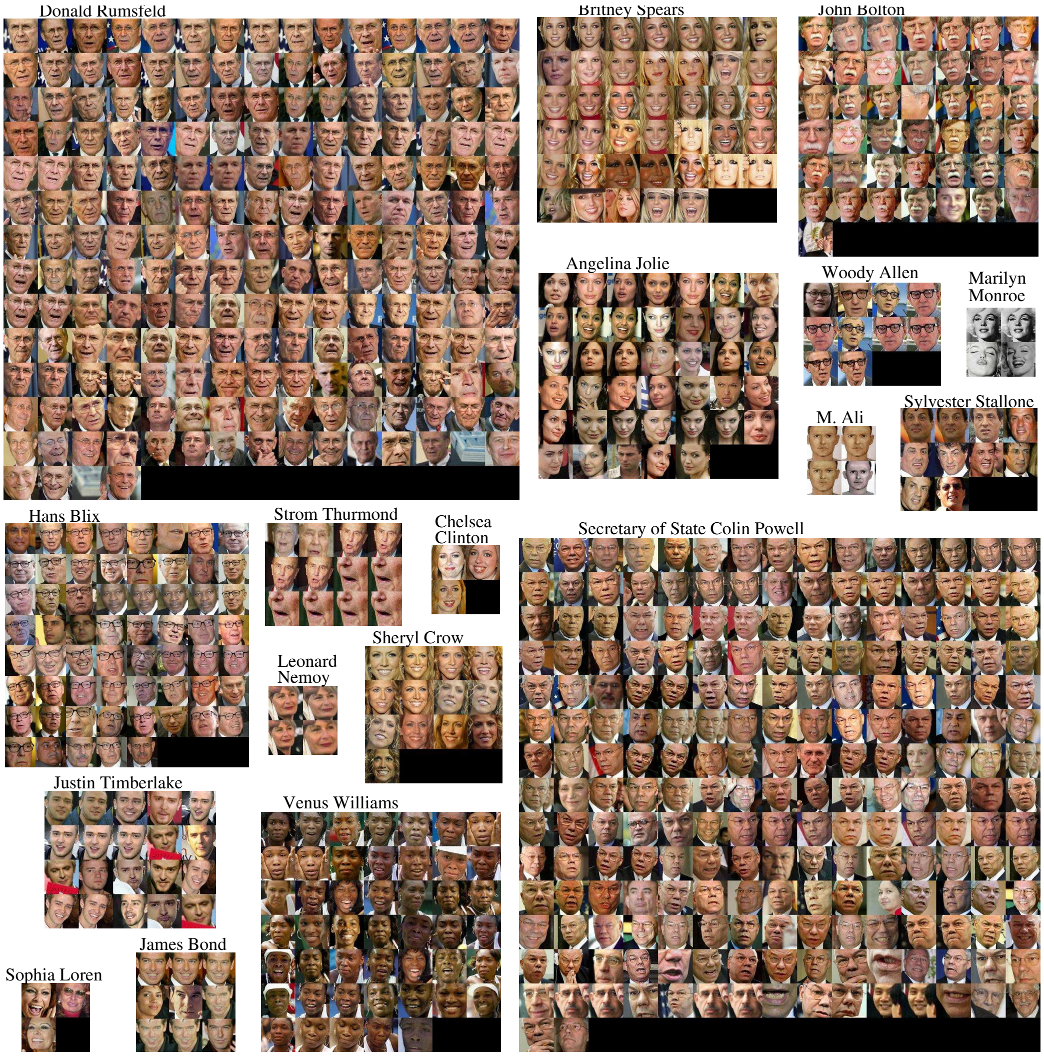


Fig. 6. *The figure shows a representative set of clusters, illustrating a series of important properties of both the dataset and the method. 1: Some faces are very frequent and appear in many different expressions and poses, with a rich range of illuminations (e.g. clusters labeled* Secretary of State Colin Powell*, or* Donald Rumsfeld*). 2: Some faces are rare, or appear in either repeated copies of one or two pictures or only slightly different pictures (e.g. cluster labeled* Chelsea Clinton *or* Sophia Loren*). 3: Some faces are not, in fact, photographs (*M. Ali*). 4: The association between proper names and face is still somewhat noisy, for example* Leonard Nemoy *which shows a name associated with the wrong face, while other clusters contain mislabeled faces (e.g.* Donald Rumsfeld *or* Angelina Jolie*). 5: Occasionally faces are incorrectly detected by the face detector (*Strom Thurmond*). 6: some names are genuinely ambiguous (*James Bond*, two different faces naturally associated with the name (the first is an actor who played James Bond, the second an actor who was a character in a James Bond film) . 7: Some faces appear in black in white (*Marilyn Monroe*) while most are in color. 8: Our clustering is quite resilient in the presence of spectacles (*Hans Blix, Woody Allen*), perhaps wigs (*John Bolton*) and mustaches (*John Bolton*).*

The Maximal Assignment procedure:

- M1 – set the δ_{ij} corresponding to the maximum likelihood assignment to 1 and all others to 0.
- M2 – maximize the parameters $P(face|name)$ using counts.

We have tried using both expectation and maximum likelihood assignment at each iteration and have found that using the maximum likelihood assignment produces better results (table 1). One possible reason for this is that in cases where there is a clear best assignment the max and the average are basically equivalent. For cases where there is no clear best, EM averages over assignments. If the probability model used by EM were the correct one, than this averaging would produce the correct results. However, in our case EM assumes a Gaussian noise model which for faces is not a great fit to the actual noise. Essentially, EM assigns too much weight to incorrect assignments, causing the expectations to be overly influenced by incorrect assignments. Using maximal assignment (in our MM procedure) avoids these issues.

4.3 Basic Clustering Evaluation

Because this is an unsupervised task, it is not meaningful to divide our data into training and test sets. Instead, to evaluate our clusterings, we create an evaluation set consisting of 1000 randomly chosen faces from our dataset. We hand label these evaluation images with their correct names (labeling with 'NULL' if the face was not named in the caption or if the named entity recognizer failed to detect the name in the caption). To evaluate a clustering, we can look at how many faces in the evaluation set are correctly labeled by that clustering.

In table 1, we see that the basic clustering correctly labels 56% of the test images correctly using EM, while the maximal assignment clustering (MM) labels 67% of the test images correctly. This clearly indicates that the maximal assignment procedure performs better than EM for our labeling task.

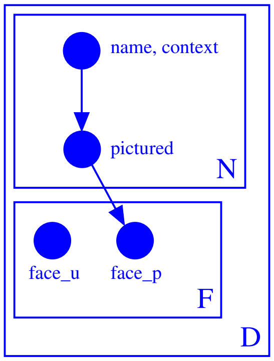
5 Clustering with Context Understanding

The context in which a name appears in a caption provides powerful cues as to whether it is depicted in the associated image. Common, quite simple phenomena in captions suggest using a language model. First, our named entity recognizer occasionally marks phrases like "United Nations" as proper names. We can determine that these names do not refer to depicted people because they appear in quite different linguistic contexts from the names of actual people. Caption writers tend to supply informative context; e.g. putting the depicted name early in the caption, using depiction indicators such as “(R)”, etc. From these linguistic cues, we can decide how likely a name is of being depicted in the associated photograph before even looking at the image.

In a caption such as “Michael Jackson responds to questioning Thursday, Nov. 14, 2002 in Santa Maria Superior Court in Santa Maria, Calif., during a $21 million lawsuit brought against him by Marcel Avram for failing to appear at two millennium concerts...”, Michael Jackson appears in a more favorable context (at the beginning of the caption, followed by a verb) than Marcel Avram (near the middle of the caption, followed by a preposition).

In the basic clustering scheme explained so far, we have ignored the context of names within the caption. By incorporating language understanding into our model we generate better assignments. Our new EM procedure uses the same procedure as our initial basic clustering method except it iterates between computing an expected set of face-name correspondences (given a face clustering and language model) and updating the face clusters and language model given the correspondences. First, we formalize our generative model of how news items are generated to incorporate a natural language model.

Our generative model is:



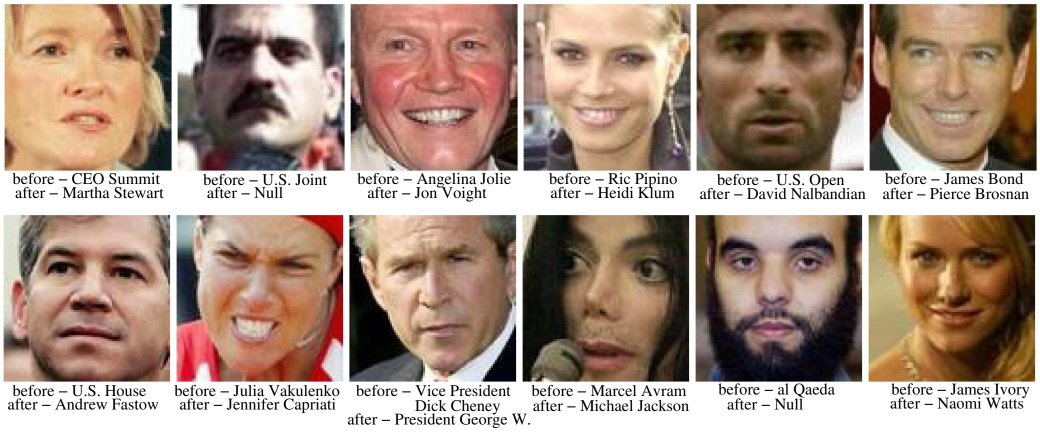
To generate a data item:

(1) Choose N, the number of names, and F, the number of faces.
(2) Generate N *name, context* pairs.
(3) For each of these *name, context* pairs, generate a binary variable $pictured$ conditioned on the context alone (from $P(pictured|context)$).
(4) For each $pictured = 1$, generate a face from $P(face|name)$.
(5) Generate $F - \sum pictured$ other faces from $P(face)$.

The parameters of this model are $P(face|name)$ (sec 3), the probability that a face, f, is generated by a name n, and $P(pictured|context)$ (sec 5.2), the probability that a name is pictured given its context.

5.1 Name Assignment

Name assignment occurs in much the same way as the basic method, but incorporates a language model that represents the probability of a name being assigned to one of the faces in the image given its context in the caption. The language model weights the names by their probability of being pictured. This allows the assign-


before – CEO Summit
after – Martha Stewart

before – U.S. Joint
after – Null

before – Angelina Jolie
after – Jon Voight

before – Ric Pipino
after – Heidi Klum

before – U.S. Open
after – David Nalbandian

before – James Bond
after – Pierce Brosnan

before – U.S. House
after – Andrew Fastow

before – Julia Vakulenko
after – Jennifer Capriati

before – Vice President Dick Cheney
after – President George W.

before – Marcel Avram
after – Michael Jackson

before – al Qaeda
after – Null

before – James Ivory
after – Naomi Watts

Fig. 7. *This figure shows some example pictures with names assigned using our raw clustering procedure* **(before)** *and assigned using a correspondence procedure with incorporated language model* **(after)**. *Our named entity recognizer sometimes detects incorrect names like "CEO Summit", but the language model assigns low probabilities to these names making their assignment unlikely. When multiple names are detected like "Julia Vakulenko" and "Jennifer Capriati", the probabilities for each name depend on their context. The caption for this picture reads "American Jennifer Capriati returns the ball to her Ukrainian opponent Julia Vakulenko in Paris during..." "Jennifer Capriati" is assigned to the face given the language model because the context in which she appears (beginning of the caption followed by a present tense verb) is more likely to be pictured than that of "Jennifer Capriati" (middle of the caption followed by a preposition). For pictures such as the one above ("al Qaeda" to "Null") where the individual is not named, the language model correctly assigns "Null" to the face. As table 1 shows, incorporating a language model improves our face clusters significantly.*

ment procedure to favor names that are more likely to be pictured based on their context over names in less probable contexts. By weighting more likely correspondences higher, we get a more accurate final assignment of faces to names.

We write P(N) as the probability of generating N names, P(F) as the probability of generating F faces, and $P(n_i, c_i)$ as the probabilities of generating $name_i$ and context c_i. For an assignment a_j, letting α index into the names that are pictured, $\sigma(\alpha)$ index into the faces assigned to the pictured names, β index into the names that are not pictured and γ index into the faces without assigned names, the likelihood of an assignment including name context is:

$$\begin{aligned} L_{x_i,a_j} =& P(N)P(F)P(n_1,c_1)...P(n_n,c_n)* \\ & \prod_{\alpha} P(pictured_{\alpha}|c_{\alpha})P(f_{\sigma(\alpha)}|n_{\alpha}) \prod_{\beta}(1 - P(pictured_{\beta}|c_{\beta})) \prod_{\gamma} P(f_{\gamma}) \end{aligned}$$

Again, the terms $P(N)P(F)P(n_1, c_1)...P(n_n, c_n)$ are not dependent on the assignment so can be ignored when calculating the probability of assignments.

The complete data log likelihood is as before:

$$\sum_{i\epsilon pics}\left[\sum_{j\epsilon C_i}(\delta_{ij}log(L(x_i,a_j))\right]$$

Where C_i are the set of possible assignments for image i, δ_{ij} is an indicator variable telling which correspondence occurred in this data item. The δ_{ij} are missing data whose expectations are computed in the E step.

This gives an EM procedure that includes updating the language models:

- E – update the δ_{ij} according to the normalized probability of picture i with assignment j.
- M – maximize the parameters $P(face|name)$ and $P(pictured|context)$ using soft counts.

5.2 Language Representation

We have explored two methods for modeling the probability of a name being pictured based on its context within a caption, $P(pictured|context)$; a Naive Bayes model in which each of the different context cues is assumed independent given the variable pictured, and a Maximum Entropy model which relaxes these independence assumptions.

5.2.1 Naive Bayes Model

For a set of context cues (C_i, for $i \in 1, 2, ...n$), our Naive Bayes model assumes that each cue is independent given the variable $pictured$. Using Bayes rule, the probability of being pictured given the cues is:

$$\begin{aligned}P(pictured|C_1,C_2,...C_n) &= \frac{P(C_1,...C_n|pictured)P(pictured)}{P(C_1,...,C_n)}\\ &= \frac{P(C_1|pictured)...P(C_n|pictured)P(pictured)}{P(C_1,...,C_n)}\\ &= \frac{P(pictured)}{P(C_1,...,C_n)}\prod_i\frac{P(pictured|C_i)P(C_i)}{P(pictured)}\\ &= \frac{1}{Z}\frac{P(pictured|C_1)...P(pictured|C_n)}{P(pictured)^{n-1}}\end{aligned}$$

Where line 1 is due to Bayes Rule, line 2 by the naive Bayes assumption, line 3 by Bayes Rule, and where the Z term in line 4 is dependent only on the cues $C_1, ..., C_n$.

Model	EM	MM
Appearance Model, No Lang Model	56%	67%
Appearance Model + N.B. Lang Model	72%	77%
Appearance Model + Max Ent Lang Model	–	78%

Table 1
Above: *To form an evaluation set, we randomly selected 1000 faces from our dataset and hand labeled them with their correct names. Here we show what percentage of those faces are correctly labeled by each of our methods (clustering without a language model, clustering with our Naive Bayes language model and clustering with our maximum entropy language model). Incorporating a language model improves our labeling accuracy significantly. Standard statistical knowledge says that EM should perform better than choosing the maximal assignment at each step. However, we have found that using the maximal assignment works better than EM for both the basic clustering and clustering with a language model. One reason this could be true is that EM is averaging faces into the mean that do not belong.*

We compute $P(pictured|C_1, ..., C_n)$ and $P(notpictured|C_1, ..., C_n)$ ignoring the Z term, and then normalize so that $P(pictured|C_1, ..., C_n)$ and $P(notpictured|C_1, ..., C_n)$ sum to 1.

We update the distributions, $P(pictured|C_i)$ and $P(pictured)$, at each iteration of our clustering process using maximum likelihood estimates based on soft counts. $P(pictured|C_i)$ is updated by of how often each context appears describing an assigned name, versus how often that context appears describing an unassigned name. $P(pictured)$ is computed using soft counts of how often names are pictured versus not pictured. We use one distribution for each possible context cue, and assume that context cues are independent when modeling these distributions (because we lack enough data to model them jointly).

For context, we use a variety of cues: the part of speech tags of the word immediately prior to the name and immediately after the name within the caption (modeled jointly), the location of the name in the caption, and the distances to the nearest ", ", ".", "(", ")", "(L)", "(R)", and "(C)" (these distances are quantized and binned into histograms). We tried adding a variety of other language model cues, but found that they did not increase assignment accuracy.

Some indications of a name being pictured learned by the Naive Bayes model were: 1. The closer the name was to the beginning of the caption, the more likely it was of being pictured, 2. The "START" tag directly before the name was a very good indicator of the name being pictured, 3. Names followed by different forms of present tense verbs were good indications of being pictured, 4. The name being followed by "(L)", "(R)" and "(C)" were also somewhat good indications of picturedness.

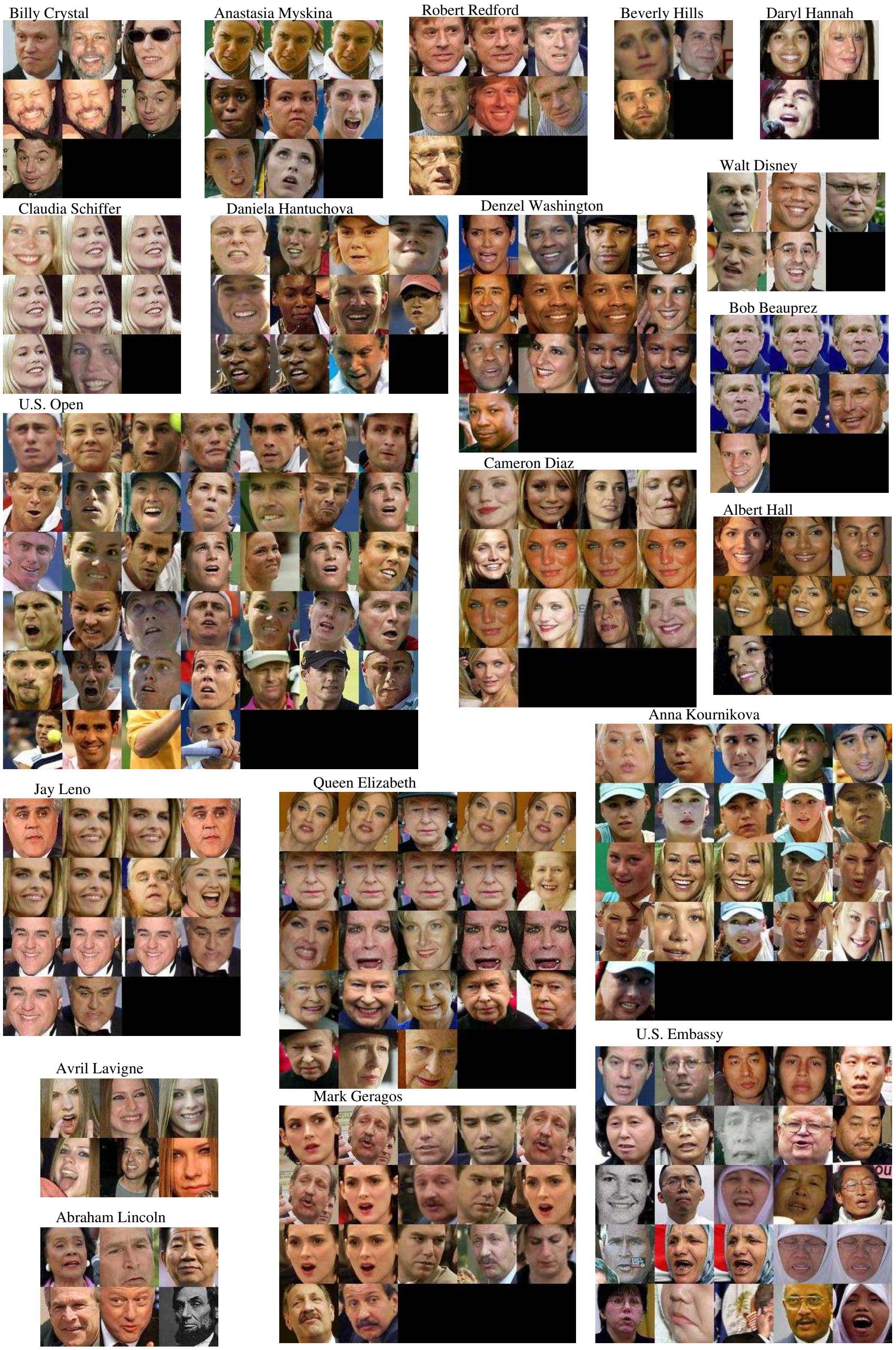


Fig. 8. *Example clusters found using our basic clustering method (see section 4 for details). Note that the names of some clusters are not actual people's names (e.g. "U.S. Open", "Walt Disney") and that there are clusters with multiple errors ("Queen Elizabeth", "Jay Leno").*

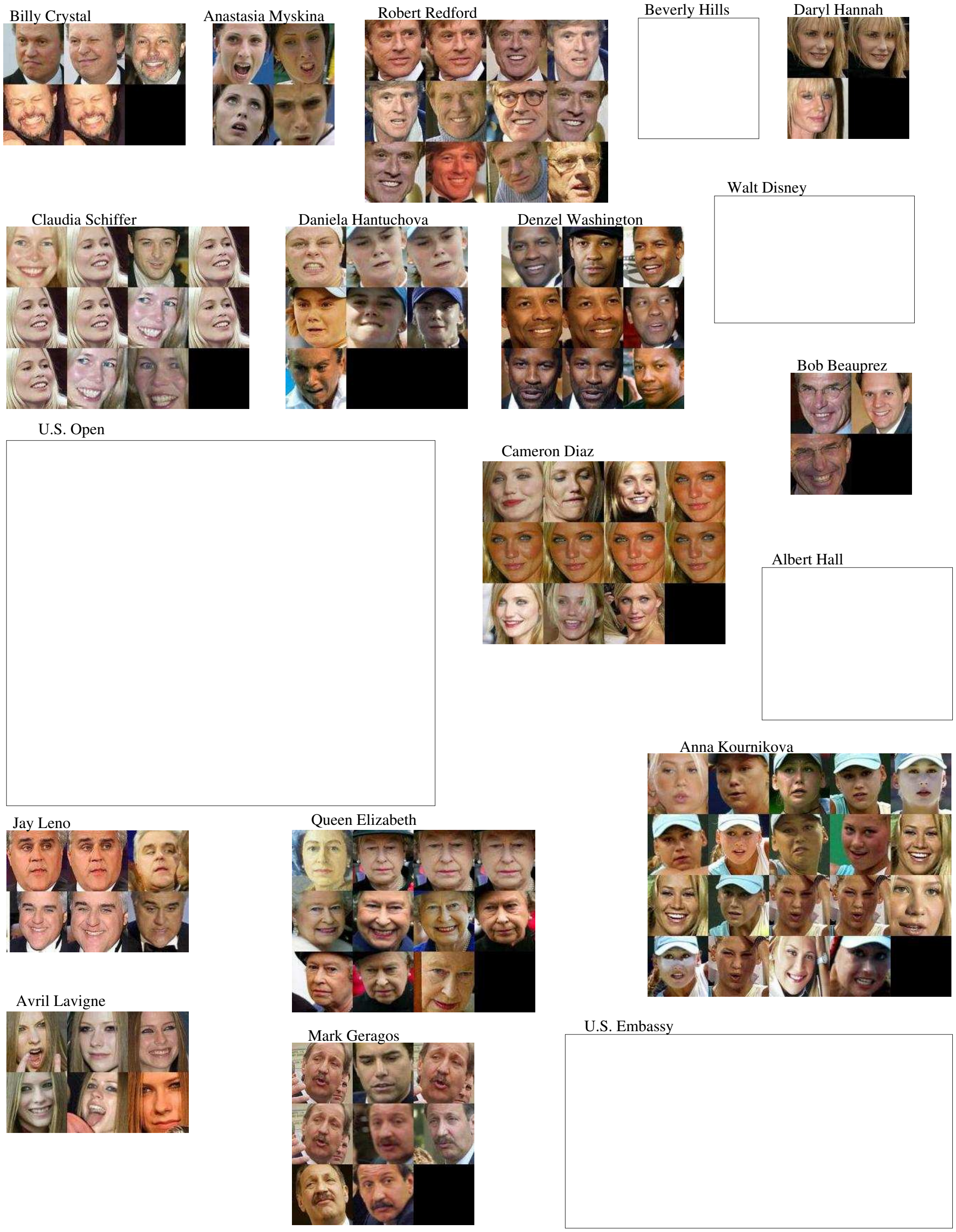


Fig. 9. *The clusters of figure 8 are improved through the use of language understanding (see section 5 for details). The context of a name within the caption often provides clues as to whether the name is depicted. By analyzing the context of detected names, our improved clustering gives the more accurate clusters seen above. The named entity recognizer occasionally marks some phrases like "U.S. Open" and "Albert Hall" as proper names. By analyzing their context within the caption, our system correctly determined that no faces should be labeled with these phrases. Incorporating language information also makes some clusters larger ("Robert Redford"), and some clusters more accurate ("Queen Elizabeth", "Bob Beauprez").*

IN Pete Sampras IN of the U.S. celebrates his victory over Denmark's **OUT Kristian Pless OUT** at the **OUT U.S. Open OUT** at Flushing Meadows August 30, 2002. Sampras won the match 6-3 7- 5 6-4. REUTERS/Kevin Lamarque

Germany's **IN Chancellor Gerhard Schroeder IN**, left, in discussion with France's **IN President Jacques Chirac IN** on the second day of the EU summit at the European Council headquarters in Brussels, Friday Oct. 25, 2002. EU leaders are to close a deal Friday on finalizing entry talks with 10 candidate countries after a surprise breakthrough agreement on Thursday between France and Germany regarding farm spending.(AP Photo/European Commission/HO)

'The Right Stuff' cast members **IN Pamela Reed IN**, (L) poses with fellow cast member **IN Veronica Cartwright IN** at the 20th anniversary of the film in Hollywood, June 9, 2003. The women played wives of astronauts in the film about early United States test pilots and the space program. The film directed by **OUT Philip Kaufman OUT**, is celebrating its 20th anniversary and is being released on DVD. REUTERS/Fred Prouser

Kraft Foods Inc., the largest U.S. food company, on July 1, 2003 said it would take steps, like capping portion sizes and providing more nutrition information, as it and other companies face growing concern and even lawsuits due to rising obesity rates. In May of this year, San Francisco attorney **OUT Stephen Joseph OUT**, shown above, sought to ban Oreo cookies in California – a suit that was withdrawn less than two weeks later. Photo by Tim Wimborne/Reuters REUTERS/Tim Wimborne

Fig. 10. *Our new procedure gives us not only better clustering results, but also a natural language classifier which can be tested separately.* **Above:** *a few captions where detected names have been labeled with IN (pictured) and OUT (not pictured) using our learned language model. Our language model has learned which contexts have high probability of referring to pictured individuals and which contexts have low probabilities. We can use this model to evaluate the context of each new detected name and label it as IN or OUT. We observe an 85% accuracy of labeling who is portrayed in a picture using only our language model. The top 3 labelings are all correct. The last incorrectly labels "Stephen Joseph" as not pictured when in fact he is the subject of the picture. Some contexts that are often incorrectly labeled are those where the name appears near the end of the caption (usually a cue that the individual named is not pictured). Some cues we could add that should improve the accuracy of our language model are the nearness of words like "shown", "pictured", or "photographed".*

5.2.2 Maximum Entropy Model

Maximum Entropy models have been used extensively in natural language systems (9) for tasks such as part of speech tagging (26), and ambiguity resolution (27). The goal of a Maximum Entropy method is to choose a model that is consistent with the observed statistics of the data, but which is otherwise as uniform as possible. To do this, we define a set of constraints based on some statistics observed in our data and choose the model that satisfies these constraints but which has maximum conditional entropy.

One attraction of Maximum Entropy models is that they give a nice way of modeling a conditional distribution with a large number of features without having to observe every combination of those features. Maximum Entropy models are also related to maximum likelihood; if we are considering distributions in the exponential family, then the maximum entropy model found will be the model in that family that maximizes the likelihood of the training data.

If y is the variable $pictured$ and x is the context of the name within the caption, then we are modeling a distribution $p(y|x)$. The context of a name consists of a binary vector (e.g. [1 0 0 0 1 0 0 ... 1]), where an element of the vector is 1 if the corresponding context cue is true and zero if it is not. We use the same cues as before except instead of binning the distance to the nearest ",", ".", "(", ")", "(L)", "(R)" and "(C)", the corresponding cue is true if the the string is within 3 words of the name. For the Maximum Entropy model we also add cues looking for specific strings ("pictured", "shown", "depicted" and "photo").

For each context cue, i, we define a set of indicator functions

$$f_i(x, y) = \begin{cases} 1 \text{ if } x(i) = 1 \text{ and } y = 0; \\ 0 \text{ otherwise.} \end{cases}$$

$$f_{2i}(x, y) = \begin{cases} 1 \text{ if } x(i) = 1 \text{ and } y = 1; \\ 0 \text{ otherwise.} \end{cases}$$

Our constraints are that the expected value of each f with respect to the training data, $\tilde{p}(f)$ is equal to the expected value of f with respect to the model p(f).

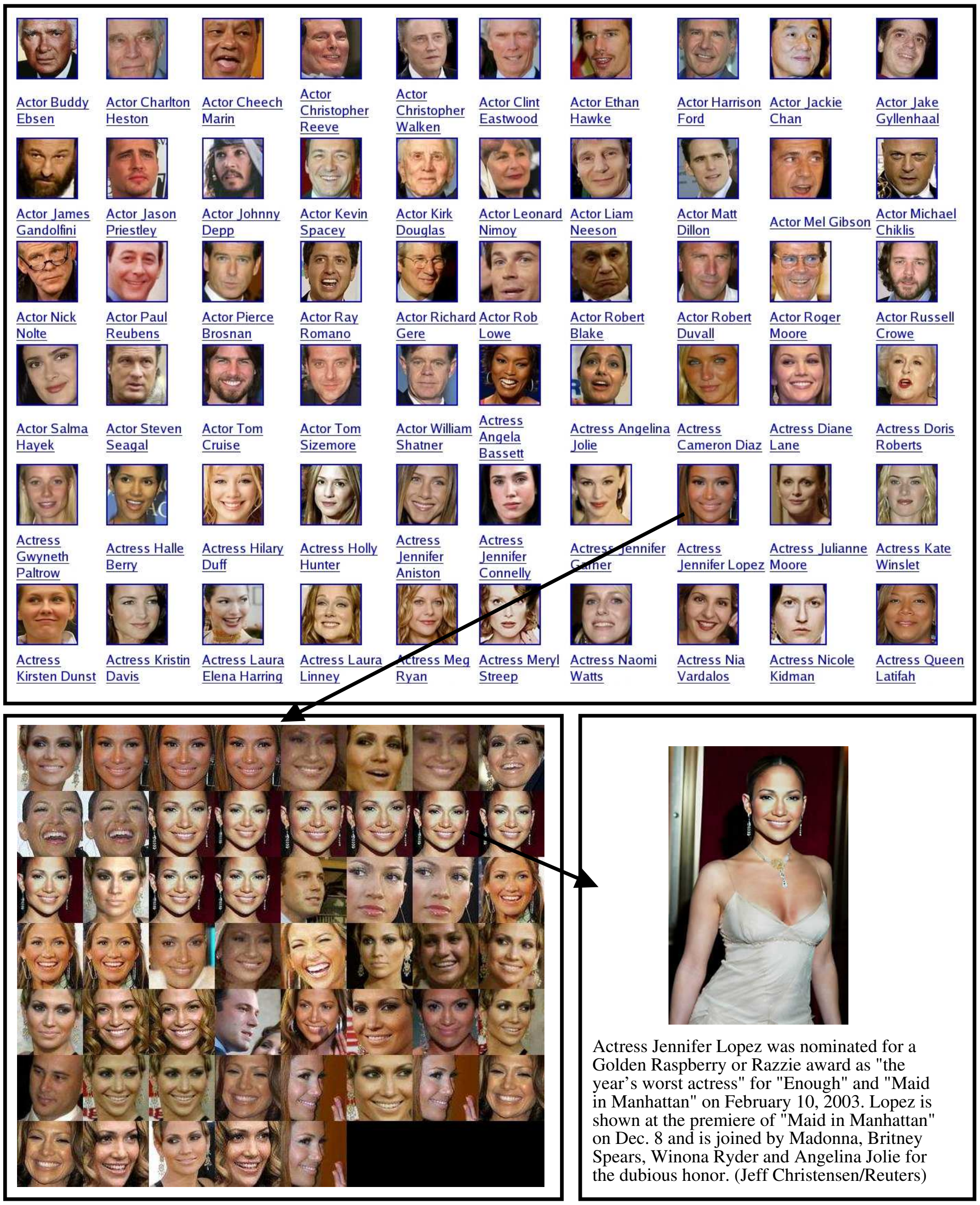
This poses an optimization problem where we want to maximize the conditional entropy of $p(y|x)$ subject to our set of constraints. If we introduce a Lagrange multiplier, λ_i for each of our constraints then we can transform this into an optimization problem where the entropy model takes the form

$$p(y|x) \propto exp \sum_i \lambda_j f_j(x, y)$$

To find the maximum likelihood $p(y|x)$, we use improved iterative scaling, the standard algorithm for finding maximum entropy distributions. Details of this model and algorithm are described in (9).

5.2.3 Comparison of language models

Using the same evaluation as section 4.3, we tested each of our language models. The models performed approximately the same on our hand labeled test set of 1000



Actress Jennifer Lopez was nominated for a Golden Raspberry or Razzie award as "the year's worst actress" for "Enough" and "Maid in Manhattan" on February 10, 2003. Lopez is shown at the premiere of "Maid in Manhattan" on Dec. 8 and is joined by Madonna, Britney Spears, Winona Ryder and Angelina Jolie for the dubious honor. (Jeff Christensen/Reuters)

Fig. 11. *We have created a web interface for organizing and browsing news photographs according to individual. Our dataset consists of 30,281 faces depicting approximately 3,000 different individuals. Here we show a screen shot of our face dictionary* **top**, *one cluster from that face dictionary (Actress Jennifer Lopez)* **bottom left** *and one of the indexed pictures with corresponding caption* **bottom right**. *This face dictionary allows a user to search for photographs of an individual as well as giving access to the original news photographs and captions featuring that individual. It also provides a new way of organizing the news, according to the individuals present in its photos.*

Classifier	labels correct	IN corr.	OUT corr.
Baseline	67%	100%	0%
EM Labeling with N.B. Language Model	76%	95%	56%
MM Labeling with N.B. Language Model	84%	87%	76%
MM Labeling with max ent Language Model	86%	91%	75%

Table 2
Above: *To form an evaluation set for text labeling, we randomly chose 430 captions from our dataset and hand labeled them with IN/OUT according to whether that name was depicted in the corresponding picture. To evaluate how well our natural language module performed on labeling depiction we look at how our test set names were labeled. "labels correct" refers to the percentage of names that were correctly labeled, "IN correct" refers to the percentage of IN names that were correctly labeled, "OUT correct" refers to the percentage of OUT names that were correctly labeled. The baseline figure gives the accuracy of labeling all names as IN. Incorporating both our Naive Bayes and Maximum Entropy language models improve labeling significantly. As with the faces, the maximum likelihood procedure performs better than EM. Names that are most often mislabeled are those that appear near the end of the caption or in contexts that most often denote people who are not pictured.*

faces. As can be seen in table 1 the Naive Bayes language model labeled 77% of the faces correctly, while the maximum entropy model labeled 78% correctly.

Another test of a language model is to see how it performs on text alone. To test this, we hand labeled the names in 430 randomly selected captions with "IN" if the name was depicted in the corresponding picture and "OUT" if it was not. On this evaluation set (without any knowledge of the associated images), the Naive Bayes model labeled 84% of the names correctly while the Maximum Entropy model labeled 86% of the names correctly (table 2). Based on these two tests, we conclude that these models perform approximately equivalently on our dataset.

5.3 Word + Face context

Given our success with linguistic context, another natural step is to incorporate context on the image side in a similar fashion to the way we used language context. For example, one might suppose that a face on the left should be given higher priority for assignment to a name that is followed by "(L)" in the associated caption. To model image context, we incorporated a maximum entropy model of face context given name context ($P(context_{face}|context_{name})$. The feature used for face context was location in the image, and for name context the features were "(L)", "(R)", "left" and "right". The maximum entropy model correctly learned that "(L)" and "left" were good indicators of the face image being on the left side of the image, while "(R)" and "right" were good indicators of the face image being on the right side of the image.

However, incorporating this model into our clustering scheme had little effect on the correctness of our labelings (only increasing the accuracy by 0.3%). The reasons this might be true are: 1. Only about 10% of all the names exhibited these context cues, 2. The names with these context cues are in general already correctly assigned by our system, and 3. The signal present in linking for example “left” and the image being on the left side of the image is fairly noisy, making their connection tentative.

6 Results

Some questions we might ask about our system are: How does analyzing language improve our face clustering results? How well does our learned natural language classifier work on text alone? How well does EM compare to using the maximal assignment procedure?

6.1 Face Classification with a Language Model

We greatly improve our face labeling accuracy by incorporating a natural language system to model name context. To evaluate how much language context helps our clustering, we use the evaluation setup described in section 4.3. Our evaluation set consists of 1000 randomly selected faces that have been hand labeled with the correct names (or ”NULL” if the person was not named in the caption or if the named entity recognizer failed to extract the correct name). On this set of faces, incorporating a language model increases our labeling accuracy from 67% without the linguistic context model, to 78% using our model of context (see table 1). We also show (table 1) that the two context models we implemented, Naive Bayes and Maximum Entropy, produce very similar results (77% and 78% respectively), implying that both models are fairly good models of name context.

By learning both a language and appearance model, we are able to achieve greater performance than using either language or vision alone. This points to the power of multi-modal data that allows us to exploit the combined information provided by each source to attain better results.

6.2 Depiction Identification in Text

One pleasing by-product of our clustering is a natural language classifier (the context model produced by our system). We can evaluate this classifier on text in isolation without an associated picture. To evaluate our context model, we create an evaluation set consisting of 430 randomly chosen captions from our dataset. We hand label these captions according to which names are depicted (“IN”) and which are not (“OUT”). By looking at how our natural language classifier labels the names within these captions, we can judge the power of our learned named entity classifier.

Figure 10 shows some example captions labeled using the leanred Maximum Entropy Context model. In table 2, we show error rates for classification of names (classified as pictured or not pictured) using our two context models, Maximum Entropy and Naive Bayes. The Maximum Entropy context model correctly labels 86% of the names, and the Naive Bayes context model 84%, while the baseline (labeling everyone as "IN") has a classification accuracy of 67%. Similarly to the face classification task, the two models perform with approximately the same accuracy, though the Maximum Entropy model again has a slight advantage over the Naive Bayes model.

6.3 EM vs MM

For missing data problems, EM is usually the preferred choice. However, we have observed that choosing the maximum likelihood assignment has a large advantage over computing an expectation. As can be seen in table 1 and table 2, the maximal assignment procedure (MM) outperforms EM in both tasks. For face labeling, MM labels 77% of the faces correctly while EM only labels 72% correctly. For text labeling, MM labels 84% of the names correctly while EM labels 76% correctly. For our task, using a hard clustering procedure (MM) has a clear advantage over the soft assignment of EM.

6.4 Initial recognition tests

We have performed several baseline recognition tests on a ground truth subset of our rectified face images consisting of 3,076 faces (241 individuals with 5 or more face images per individual). The cluster of faces for each individual were used and hand cleaned to remove erroneously labeled faces. Half of the individuals were used for training, and half for testing. Two common baselines for face recognition datasets are PCA and PCA followed by LDA. On the test portion of this set, using the first 100 basis vectors found by PCA on the cropped face region with a 1-Nearest Neighbor Classifier gives recognition rates: of 9.4% ± 1.1% using a gallery set of one face per individual, 12.4% ± 0.6% using a gallery of two faces per individual, and 15.4% ± 1.1% using a gallery set of three faces per individual.

Using the first 50 basis vectors of LDA computed on the PCA vectors increases the accuracy to: 17% ± 2.4% for a gallery of one face per individual, 23% ± 1.9% for a gallery of two faces per individual and 27.4% ± 2.6% for a gallery of 3 faces per individual. These numbers are quite a bit lower than the 80-90% baseline recognition rates quoted for most datasets, suggesting that our face images are in fact quite challenging and that they will be a useful dataset for training future face recognition systems.

7 Conclusion

We have automatically produced a very large and realistic face dataset consisting of 30,281 faces with roughly 3,000 different individuals from news photographs with associated captions. This dataset can be used for further exploration of face recognition algorithms. Using simple models for images and text, we are able to create a fairly good assignment of names to faces in our dataset. By incorporating contextual information, this labeling is substantially improved, demonstrating that words and pictures can be used in tandem to produce results that are better than using either medium alone.

Another product of our system is a web interface that organizes the news in a novel way, according to individuals present in news photographs. Users are able to browse the news according to individual (Figure 11), bring up multiple photographs of a person and view the original news photographs and associated captions featuring that person.

We can use the language and appearance models learned by our system to label novel images or text in isolation. By learning these models in concert, we boost the amount of information available from either the images and text alone. This increases the performance power of our learned models. We have conclusively shown that by incorporating language information we can improve a vision task, namely automatic labeling of faces in images.

References

[1] O. Arandjelovic, A. Zisserman. “Automatic Face Recognition for Film Character Retrieval in Feature-Length Films”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[2] F.R. Bach, M.I. Jordan, “Kernel independent component analysis”, *International Conference on Acoustics, Speech, and Signal Processing*, 2003

[3] K. Barnard,P. Duygulu, N. de Freitas, D.A. Forsyth, D. Blei, M.I. Jordan, “Matching Words and Pictures”, *Journal of Machine Learning Research*, Vol 3, pp. 1107-1135, 2003.

[4] K. Barnard and P. Duygulu and D.A. Forsyth, “Clustering Art”, *Computer Vision and Pattern Recognition*, Vol II, pp. 434-441, 2001.

[5] P. Belhumeur, J. Hespanha, D. Kriegman “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection” *Transactions on Pattern Analysis and Machine Intelligence*, Special issue on face recognition, pp. 711-720, July 1997.

[6] A.C. Berg, J. Malik, “Geometric Blur for Template Matching,” *Computer Vision and Pattern Recognition*,Vol I, pp. 607-614, 2001.

[7] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, E. Learned-Miller, D.A. Forsyth “Names and Faces in the News” *Computer Vision and Pattern Recognition*, 2004.

[8] T.L. Berg, A.C. Berg, J. Edwards, D.A. Forsyth “Who’s in the Picture” *Neural Information Processing Systems*, 2004.

[9] A. Berger, S.D. Pietra, V. D. Pietra, “A maximum entropy approach to natural language processing”, *Computational Linguistics*, Vol. 22-1, March 1996.

[10] V. Blanz, T. Vetter, “Face Recognition Based on Fitting a 3D Morphable Model,” *Transactions on Pattern Analysis and Machine Intelligence* Vol. 25 no.9, 2003.

[11] C. Carson, S. Belongie, H. Greenspan, J. Malik, “Blobworld – Image segmentation using expectationmaximization and its application to image querying,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8), pp. 1026–1038, 2002.

[12] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, “GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications,” *40th Anniversary Meeting of the Association for Computational Linguistics”*, Philadelphia, July 2002.

[13] P. Duygulu, K. Barnard, N. de Freitas, D.A. Forsyth “Object Recognition as Machine Translation”, *European Conference on Computer Vision*, Vol IV, pp. 97-112, 2002.

[14] J. Edwards, R. White, D.A. Forsyth, “Words and Pictures in the News,” *Workshop on Learning Word Meaning from Non-Linguistic Data*, 2003.

[15] M. A. Fischler and R.C. Bolles, “Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography”, *Comm. ACM*, Vol.24, pp. 381-395, 1981.

[16] A.W. Fitzgibbon, A. Zisserman: “On Affine Invariant Clustering and Automatic Cast Listing in Movies,” *European Conference on Computer Vision*, 2002

[17] C. Fowlkes, S. Belongie, F. Chung and J. Malik, “Spectral Grouping Using The Nyström Method,” *TPAMI*, Vol. 26, No. 2, February 2004.

[18] R. Gross, J. Shi and J. Cohn, “Quo Vadis Face Recognition?,” *Third Workshop on Empirical Evaluation Methods in Computer Vision*, December, 2001.

[19] R. Gross, I. Matthews, and S. Baker, “Appearance-Based Face Recognition and Light-Fields,” *Transactions on Pattern Analysis and Machine Intelligence*, 2004.

[20] V. Lavrenko, R. Manmatha., J. Jeon, “A Model for Learning the Semantics of Pictures,” *Neural Information Processing Systems*, 2003

[21] T. Leung, M.C. Burl, and P. Perona, “Finding Faces in Cluttered Scenes using Random Labelled Graph Matching”, *Int. Conf Computer Vision*, 1995.

[22] J. Li and J. Z. Wang, “Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach,” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1075-1088, 2003

[23] K. Mikolajczyk “Face detector,” *Ph.D report*, INRIA Rhone-Alpes

[24] P.J. Phillips, P. Grother, R.J Micheals, D.M. Blackburn, E Tabassi, J.M. Bone, “FRVT 2002: Evaluation Report”, *Technical report, Face Recognition Vendor Test*, 2002.

[25] P. J. Phillips, E. Newton, “Meta-analysis of face recognition algorithms”, *Int. Conf.*

on Automatic Face and Gesture Recognition, 2002.

[26] A. Ratnaparkhi, “A Maximum Entropy Model for Part-Of-Speech Tagging” *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1996.

[27] A. Ratnaparkhi, “Maximum Entropy Models For Natural Language Ambiguity Resolution” *PhD Thesis*, 1998.

[28] S. Romdhani, V. Blanz, T. Vetter, “Face identification across different poses and illumination with a 3d morphable model”, *International Conference on Automatic Face and Gesture Recognition*, 2002.

[29] J. Scheeres, “Airport face scanner failed”, *Wired News*, 2002. http://www.wired.com/news/privacy/0,1848,52563,00.html.

[30] C. Schmid, “Constructing models for content-based image retrieval”, *Computer Vision and Pattern Recognition*, 2001.

[31] L. Sirovitch, M. Kirby, “Low-dimensional procedure for the characterization of human faces”, *J. Opt. Soc. Am.* Vol 2, pp. 586-591, 1987.

[32] H. Schneiderman, T. Kanade, “A statistical metho d for 3D object detection applied to faces and cars”, *Computer Vision and Pattern Recognition*, Vol. I, pp. 746-751, 2000.

[33] B. Scholkopf, A. Smola, K.-R. Muller “Nonlinear Component Analysis as a Kernel Eigenvalue Problem”, *Neural Computation*, Vol. 10, pp. 1299-1319, 1998.

[34] P.H.S. Torr, D.W. Murray, “The development and comparison of robust methods for estimating the fundamental matrix”, *Internation Journal on Computer Vision*, Vol. 24, pp. 271-300, 1997.

[35] M. Turk, A. Pentland, “Face Recognition using Eigenfaces”, *Computer Vision and Pattern Recognition*, pp. 586-591, 1991.

[36] P.H.S. Torr, A. Zisserman, “MLESAC: A New Robust Estimator with Application to Estimating Image Geometry”, *Computer Vision and Image Understanding*, Vol. 78, pp. 138-156, 2000.

[37] P. Viola, M. Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features”, *Computer Vision and Pattern Recognition*, 2001,

[38] C. Williams, M. Seeger “Using the Nyström Method to Speed up Kernel Machines”, *Advances in Neural Information Processing Systems*, Vol 13, pp. 682-688, 2001.

[39] M.H. Yang, N. Ahuja, D. Kriegman, “ Face Recognition Using Kernel Eigenfaces”, *Int. Conf. on Image Processing*, vol. 1, pp. 37-40, 2000

[40] M.H. Yang, D. Kriegman, N. Ahuja “Detecting Faces in Images : A Survey”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24-1, pp.34-58, 2002.

[41] W. Zhao, R. Chellapa, and A. Krishnaswamy, “Discriminant analysis of principal components for face recognition,” *International Conference on Automatic Face and Gesture Recognition*, pp. 336-341, 1998.