Names and Faces

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

Abstract

We show that a large and realistic face data set can be built from news photographs and their associated captions. Our automatically constructed face data set consists of 30,281 face images, obtained by applying a face finder to approximately half a million captioned news images. The faces are labeled using image information from the photographs and word information extracted from the corresponding caption. This data set is more realistic than usual face recognition data sets, because it contains faces captured "in the wild" under a wide range of positions, poses, facial expressions, and illuminations. After 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 their associated names in the 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.

keywords: Names; Faces; News; Words; Pictures

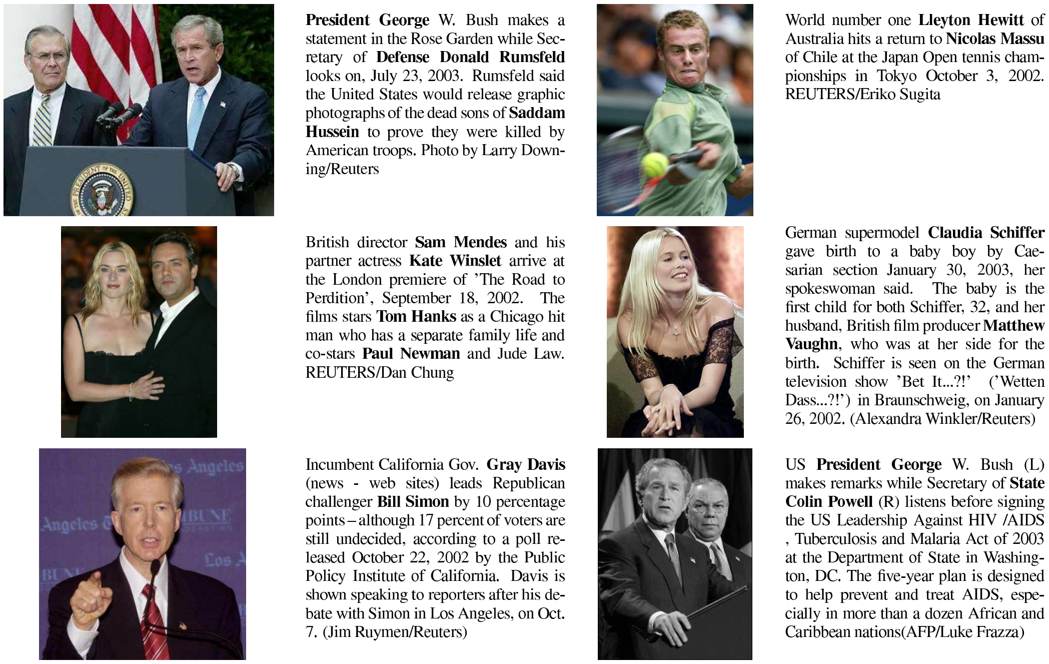
1 Introduction

This paper shows how to exploit the success of face detection to build a rich and reasonably accurate collection of labeled faces. The input is a collection of news photographs with captions. Face detection extracts faces from each image while natural language processing finds proper names in the associated caption. For each photo/caption pair, a *data item*, the remaining step, to solve the assignment problem between names and faces, is the central part of this article.

We attack the assignment problem in two ways. First we develop an iterative method for determining correspondences for a large number of data items, along a familiar line of reasoning. If we knew an appearance model for the faces associated with each name, then finding a correspondence would be straightforward; similarly if we knew a correspondence then estimating an appearance model for the faces associated with each name would be straightforward. These observations lead to natural iterative algorithms. Second we show that there are contextual language cues that suggest particular names in a

---

*CS Division, U.C. Berkeley
†Gatsby Computational Neuroscience Unit, UCL
‡CS Department, U.Mass Amherst
§CS Department, UIUC



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)

Figure 1: *Some typical news photographs with associated captions from our data set. Notice that multiple faces may appear in a single picture and that multiple names may occur in a particular caption. 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). The output of our system on these images appears in Figure 5.*

caption do not refer to a pictured face. These cues are learned and exploited in the iterative algorithms, improving the resulting correspondences.

1.1 Previous work

There are many data sets of images with associated words. Examples include: collections of museum material [3]; the Corel collection of images ([4, 22, 17], and numerous others); any video with sound or closed captioning [57, 56, 71]; images collected from the web with their enclosing web pages [11]; or captioned news images [70]. It is a remarkable fact that, in these collections, pictures and their associated annotations are complementary. The literature is very extensive, and we can mention only the most relevant papers here. For a more complete review, we refer readers to [19], which has 120 references. There are three natural activities: One might wish to cluster images, to search for images using keywords, or to attach keywords to new images. Typically, models intended for one purpose can produce results for others.

Search: Belongie *et al.* demonstrate examples of joint image-keyword searches [16]. Joshi *et al.* show that one can identify pictures that illustrate a story by searching annotated images for those with relevant keywords, then ranking the pool of images based on similarity of appearance [36].

Clustering: Barnard *et al.* cluster Corel images and their keywords jointly to produce a browsable representation [4]; the clustering method is due to Hofmann and Puzicha [31]. Barnard *et al.* show that this form of clustering can produce a useful, browsable representation of a large collection of annotated art in digital form [3].

Attaching keywords to images: Clustering methods methods can typically be used to predict keywords from images, and accuracy at keyword prediction is used as one test of such methods (see also [6]). There are two varieties of the prediction task: predicting words associated with an image (*auto-annotation*) and predicting words associated with particular image structures. Maron and Ratan attach keywords to images using *multiple-instance learning* [42]. Multiple-instance learning is a general strategy to build classifiers from "bags" of labeled examples. Typically, one knows only that a bag contains or does not contain a positive example, but not which example is positive. Methods attempt to find small regions in the feature space that appear in all positive bags and no negative bags; one can visualize these methods either as a form of smoothing [43, 82], fitting an SVM [1, 65], or using geometric reasoning [20]. Comparisons between methods appear in [51]. Chen and Wang describe a variant multiple-instance learning method, and use it to predict keywords from regions [17]. Duygulu *et al.* use explicit correspondence reasoning to associate keywords with image regions [22], using a statistical translation model from [15]. Blei and Jordan use a variant of latent Dirichlet allocation to predict words corresponding to particular image regions in an auto-annotation task [14]. Barnard *et al.* demonstrate and compare a wide variety of methods to predict keywords, including several strategies for reasoning about correspondence directly [5]. Li and Wang used 2-dimensional multi-resolution hidden markov models on categorized images to train models representing a set of concepts [39]. They then used these concepts for automatic linguistic indexing of pictures. Jeon *et al.* demonstrate annotation and retrieval with a cross-media relevance model [35]. Lavrenko *et al.* used continuous space relevance models to predict the probability of generating a word given image regions for automatic image annotation and retrieval [38].

Other activities: Relations between text and images appear to be deep and complex. Barnard and Johnson show one can disambiguate the senses of annotating words using image information [7]. Berg and Forsyth show that one can find images of complex categories ("monkey"; "penguin") by searching for images with distinctive words nearby and containing distinctive image structures [11]. Yanai and Barnard use region entropy to identify words that have straightforwardly observed visual properties ("pink" does, "affectionate" does not) [73]. All this work has tended to emphasize general image constructs (such as regions), but one might instead use detectors and link the detector responses with words. Faces are of particular interest.

1.1.1 Face Recognition

We review only important points, referring readers to Zhao *et al.* for a comprehensive general survey of the area [85]. Further reviews appear in [29, 78, 49]. Early work uses nearest neighbor classifiers based on pixel values, typically dimensionality reduced using principal component analysis (PCA) [61, 68]. Linear discriminant methods offer an improvement in performance [8]. More recently, it has been shown that models based on 3D structure, lighting, and surface appearance [13, 49] or appearance based methods that explicitly model pose [28] give better recognition accuracy, but can be somewhat hard to fit for arbitrary faces.

Face recognition is known to be difficult, and applications have failed publicly [58]. Philips and Newton show that the performance of a face recognition system on a data set can largely be predicted by the performance of a baseline algorithm, such as principal component analysis, on the same data set [48]. Since recognition systems work well on current face data sets, but poorly in practice, this suggests that the data sets currently used are not representative of real world settings. Because current data sets 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 data set that is more realistic, with wide variations in color, lighting, expression, hairstyle and elapsed time.

1.1.2 Linking Faces with Other Data

It appears to be considerably simpler to choose one of a few names to go with a face than it is to identify the face. This means one might be able to link faces with names in real data sets quite successfully. Very good face detectors are now available (important samples of this huge literature include [50, 52, 53, 54, 55, 64, 33, 69, 78, 59, 45]); we use the detector of [45]). Attempts to link names and faces appear quite early in the literature. Govindaraju *et al.* describe a method that finds faces using an edge curve criterion, and then links faces to names in captions by reasoning about explicit relational information in the caption (they give the example of the caption "Cardinal O'Connor (center), George Bush (left) and Michael Dukakis...") [27]. There is a description of an expanded version of this system, which uses language semantics even more aggressively (for example, the system possesses the knowledge that a portrait is a face surrounded by a frame, p. 53), in [63]. Zhang *et al.* show that a text based search for an image of a named individual is significantly improved by testing to see whether returned images contain faces [83]. Naaman *et al.* show that labels used frequently for "nearby" images constrain the labels that can be used for the current face image [46].

Satoh and Kanade work with video and a transcription [57]. They represent faces using principal components, identify named entities in the transcript, and then build a smoothed association between principal component faces and names that appear nearby in the transcript. Similar faces appearing near two instances of a name reinforce the association function; different names appearing near similar faces weaken it. The system operates on some 320 face instances (taken from some 4.5 hours of video) and 251 name instances, and reports names strongly associated with a given face. Satoh *et al.* describe a variant of this system, which is also capable of reading captions overlaid on video frames; these prove to be a strong cue to the identity of a face [56]. The method is comparable with multiple-instance learning methods (above). Yang and Hauptmann describe a system for learning such association functions [75].

Houghton works with video, transcriptions, automatically interpreted video captions and web pages (from news and other sources), to build a database of named faces [32]. The question of correspondence is not addressed; the data appears to contain only single face/single name pairs. Houghton's system will produce an N-best list of names for query faces.

An important nuisance in news video are anchor persons, whose faces appear often and are often associated with numerous names. Song *et al.* detect and remove anchor persons and then use a form of multiple-instance learning to build models of two well-known individuals from video data [62].

Yang *et al.* compare several forms of multiple-instance learning for attaching one of a set of possible labels to each face image [76]. In their problem, each image has a set of possible name labels, and one knows whether the right label appears in that set (there are 234 such images) or not (242). There are approximately 4.7 available labels for each face image. The paper compares four multiple-instance algorithms, each in two variants (one either averages over correspondences between a face and labels, or chooses the best correspondence) and each with two types of training data (only positive bags vs. all bags), and two supervised methods. Multiple-instance methods label between 44% and 60% of test images correctly and supervised methods label between 61% and 63% of test images correctly.

Methods to label faces in consumer images are described in [80, 81]. In this problem, the user acts as an oracle — so there is no correspondence component — but the oracle must not be queried too often.

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)

Figure 2: *In our initial set of photo-caption pairs, some individuals, like President Bush (***right***), appear frequently. Most people, however, like Dr. Nikola (***left***), appear only a few times or in only one picture. This distribution reflects what we 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 data sets with these features is necessary for producing reliable face recognition systems.*

Fitzgibbon and Zisserman automatically discover cast listings in video using affine-invariant clustering methods on detected faces and are robust to changes in lighting, viewpoint and pose [24]. More recently, Arandjelovic and Zisserman have extended this work to suppress effects of background surrounding the face, refine registration and allow for partial occlusion and expression change [2].

Our efforts differ from the work surveyed above in three important points. First, our typical data item consists of representations of several faces *and* of several names, and we must identify what, if any, correspondences are appropriate. Second, we reason explicitly about correspondence. This allows us to build discriminative models that can identify language cues that are helpful. Third, we operate at a much larger scale (approximately 30,000 face images), which can help to make correspondence reasoning more powerful.

1.2 Overview

We have collected a very large data set of captioned news images (section 2). We describe our construction of a face dictionary as a sequence of three steps. First, we detect names in captions using an open source named entity recognizer [18]. Next, we detect and represent faces, as described in section 3.3. Finally, we associate names with faces, using either a clustering method (section 4) or an enhanced method that analyzes text cues (section 5).

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. There appear to be significant benefits in explicit correspondence reasoning, and we report results for name-face association that are a significant improvement on those of Yang *et al.* [76] described above.

The result is a labeled data set of faces, captured "in the wild." This data set displays a rich variety of phenomena found in real world face recognition tasks — significant variations in color, hairstyle, expression, etc. Equally interesting is that it does *not* contain large numbers of faces in highly unusual and seldom seen poses, such as upside down. Rather than building a database of face images by choosing arbitrary ranges of pose, lighting, expression and so on, we simply let the properties of a "natural" data source determine these parameters. We believe that in the long run, developing detectors, recognizers, and other computer vision tools around such a database will produce programs that work better in realistic everyday settings.

While perfect automatic labeling is not yet possible, this data set has already proven useful, because it is large, because it contains challenging phenomena, and because correcting labels for a subset is relatively straightforward. For example, Ozkan and Duygulu [47] used the most frequent 23 people in the database (each of whom occurred over 200 times) in additional clustering experiments. The assumption made in that work is that the clusters were more than 50 percent correctly labeled, so the current data set easily met this requirement.

The database was also used recently in work by Ferencz *et al.* on face recognition [23, 34]. In this case, a subset of images with correct labels were used for training and testing in a supervised learning framework. A simple interface was used in which database examples could quickly be manually classified as correct or incorrect. As a result, it took just a couple of hours to produce a database of more than 1000 correctly labeled set of "faces in the wild" for that work.

2 News Data Set

We have collected a data set 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 [45], we extract faces from these images. Using Cunningham *et al.*'s open source named entity recognizer [18], we detect proper names in each of the associated captions. This gives us a set of faces and names resulting from each captioned picture. In each picture-caption pair, There may be several faces and several names. Furthermore, some faces may not correspond to any name, and some names may not correspond to any face. Our task is to assign one of these names or null (unnamed) to each detected face.

This collection differs from typical face recognition data sets in a number of important ways:

- **Pose, expression and illumination** vary widely. We often encounter the same face illuminated with markedly different colored light and in a very broad range of expressions. The parameters of the camera or post-processing add additional variability to the coloring of the photos. Spectacles and mustaches are common (Figure 5.4). There are wigs, images of faces on posters, differences in resolution and identikit pictures (e.g. Figure 5.4). 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, adding an additional source of variability that has been shown to hamper face recognition substantially [29].

- **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 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.

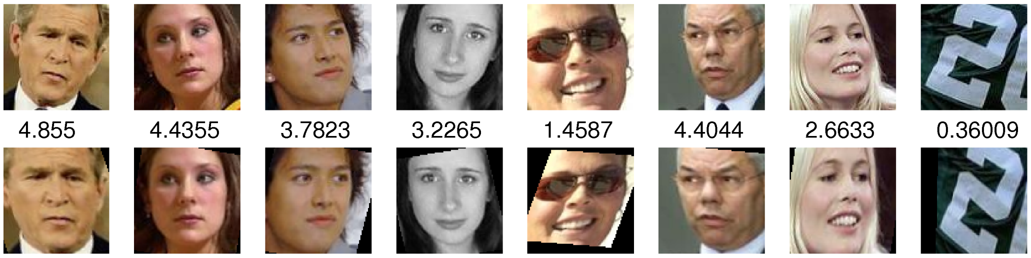


Figure 3: *The face detector can detect faces in a range of orientations, as the* **top row** *shows. Before clustering the face images we rectify them to a canonical pose* **bottom row**. *The faces are rectified using a set of SVM's trained to detect feature points on each face. Using gradient descent 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.*

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 data set 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).

One important difficulty is that our face detector cannot detect lateral or three-quarter views. This is a general difficulty with face detectors (all current face detectors either can detect only frontal views, or are significantly less reliable for views that are not frontal [78]), but it means that our data set contains only frontal or near-frontal views. We speculate that methods like ours could be made to work to produce a similar data set if one had a face detector that was aspect insensitive, but do not know what performance penalty there would be. For extremely large data sets, we expect that there may be little penalty. This is because, in a sufficiently large data set, we might reasonably expect to see many aspects of each individual in contexts where there is little ambiguity. For smaller data sets the problem would be much more challenging and would require more sophisticated representations.

3 Finding and Representing Faces

To deal with the large quantity of data, we establish a pipeline that takes in images and outputs a description based on a rough alignment of facial features. Subsequently, we compare faces in this domain.

Our pipeline is as follows. 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). We discard faces where the rectifier cannot find good base points, resulting in 34,623 faces.
3. Identify faces with at least one proper name identified in the associated caption, leaving 30, 281 faces.
4. Transform this set of faces into a representation suitable for the assignment task (Section 3.3).

3.1 Face detection

For face detection, we use Mikolajczyk's implementation [45] of the face detector described by Schneiderman and Kanade [59]. 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; a probability model then determines whether the image is a face image or a non-face image. We threshold on face size (86x86 pixels or larger) and detection score to obtain 44,773 face images.

3.2 Rectification

The next stage in our pipeline is an alignment step. While the detector detects only frontal or near frontal faces, these faces are still subject to small out of plane rotations and significant in-plane rotations and scales. We will use an appearance feature to compare face images, and so would like to reduce within-class variance and increase between-class variance. Within-class variance in appearance features can be significantly reduced by moving each face image to a canonical frame (where eyes, nose, mouth, etc. lie close to canonical locations), a procedure we call *rectification*. We will rectify by using a novel procedure to identify a set of base points in the image, then apply a full plane affine transformation to move these base points to canonical locations. Images where base points can not be identified sufficiently well will be rejected.

Notice that rectification could suppress features that help identify individuals. For example, some individuals have larger faces than others do, and rectification suppresses this property, thereby reducing between-class variance. In this application, the suppression of within-class variance obtained by rectifying faces seems to outweigh the loss of between class variance. We speculate that in a sufficiently large data set, rectification may be unnecessary, because one would have enough examples of any individual's face in any view; we have no reason to believe our data set is anywhere large enough for this to apply.

3.2.1 Identifying Base Point Locations

We train five support vector machines (SVMs) 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 feature of Berg et al [9, 10] applied to gray-scale patches as the features for our SVM.

The geometric blur descriptor first produces sparse channels from the grey scale image. In this case, these are 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 significant gains in rectification accuracy. Using geometric blur features instead of raw image patches was a necessary step to making our rectification system effective.

We compute the output value for each SVM at each point in the entire image and multiply with a weak prior on location for each feature. This produces a set of five feature maps, one for each base point. The initial location of each base point is obtained as the maximal point of each map.

3.2.2 Computing the Rectification

We compute an initial affine map from canonical feature locations to the initial locations for the base points using least squares. However, accepting a small decrease in the SVM response for one base point may be rewarded by a large increase in the response for another. We therefore maximize the sum of SVM responses at mapped canonical feature locations using gradient descent, with the initial affine map as a start point. The image is then rectified using the resulting map, and the value of the optimization problem is used as a score of the rectification. The value indicates how successful we have been at finding base points; a small score suggests that there is no set of points in the image that (a) looks like the relevant face features and (b) lies near to the result of an affine map applied to the canonical points.

We filter our data set by removing images with poor rectification scores, leaving 34,623 face images. This tends to remove the face detectors false positives (Figure 2; center number – larger numbers indicate a better score). Each face is then automatically cropped to a region surrounding the eyes, nose and mouth in the canonical frame, to eliminate effects of background on recognition. The RGB pixel values from each cropped face are concatenated into a vector and used as a base representation from here on.

3.3 Face Representation

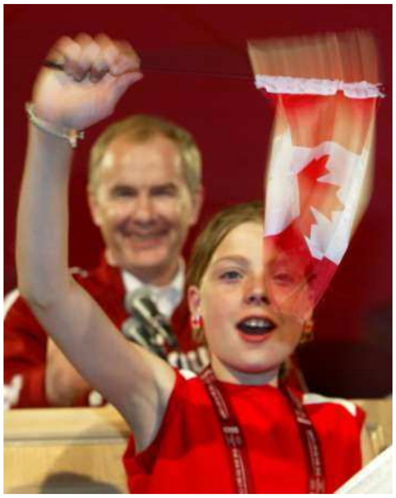
We wish to represent faces appearances as vectors in a space where, if one uses Euclidean distance between vectors, examples of the same face are close together and examples of different faces are far apart. We must identify components of the base representation that tend to be similar for all faces, and discard them (or, better, keep components of the base variation that vary strongly over the data set). Of these, we must keep those that tend to co-vary with identity.

We use kernel principal component analysis (kPCA; see [60]) to identify components of the base representation that vary strongly over the data set. The result is a vector of kernel principal components. We apply linear discriminant analysis (LDA; see, for example [30]) to these vectors, to obtain a feature vector. Kernel principal component analysis is a standard method of dimension reduction that has been shown to be effective for face recognition (see, for example [40, 37, 77, 74, 41, 84]; Yang compares with principal components and with linear discriminant analysis and shows a strong advantage for kPCA combined with LDA [79]).

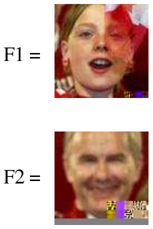
3.3.1 Kernel Principal Components and the Nyström Approximation

Kernel Principal Components Analysis requires the following steps:

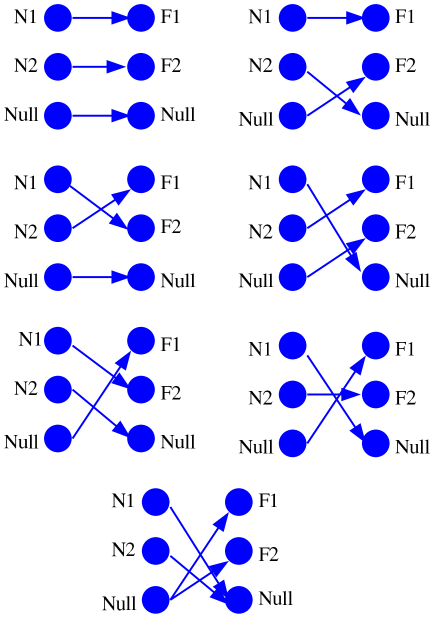
- Compute a kernel matrix, K, where $K_{ij} = K(\text{image}_i, \text{image}_j)$ is the value of a kernel function comparing image_i and image_j. We use a Gaussian kernel with sigma set to produce reasonable



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



N1 = John Furlong,
N2 = Winter Olympics,



All possible name/face assignments

Figure 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".*

kernel values.

- 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.

Writing N for the number of data items, we have an NxN kernel matrix. In our case, $N = 34,623$, and we cannot expect to evaluate every entry of the matrix. We do not use incomplete Cholesky decomposition, which can give a bound on the approximation error [26], because that would require accessing all images for each column computation. However, the kernel matrix must have relatively low column rank; if it did not, there would be generalization problems, because it would be difficult to predict a column from the other columns (see [72]). This suggests using the Nyström approximation method, which will be accurate if the matrix does have low column rank, and which allows the images to be accessed only once in a single batch rather than once for each column computation (cf. [72, 25]).

The Nyström method partitions the kernel matrix 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 $C \in \mathbb{R}^{(N-n) \times (N-n)}$. To obtain A and B by selecting a base set $\mathcal{B}$ of the set of images $\mathcal{I}$ (in our case, 1000 images selected uniformly and at random). Then

$$A_{uv} = K(\text{image}_u, \text{image}_v) \text{ for image}_u \in \mathcal{B}, \text{image}_v \in \mathcal{B},$$

where $K(\cdot, \cdot)$ is the kernel function, and

$$B_{lm} = K(\text{image}_l, \text{image}_m) \text{ for image}_l \in \mathcal{B}, \text{image}_m \in \mathcal{I}.$$

Now Nyström's method approximates K with the matrix obtained by replacing C with $\hat{C} = B^T A^{-1} B$, yielding $\hat{K} = \begin{bmatrix} A & B \\ B^T & \hat{C} \end{bmatrix}$.

Centering: we center $\hat{K}$ as usual in kPCA, by writing 1_N for an $Nx1$ vector of ones, and then computing

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

Note that this is simplified by the fact that $\hat{K}$ is symmetric, and by observing that

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

It is convenient to write $\tilde{K} = \begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{B}^T & \tilde{C} \end{bmatrix}$ where the dimensions of $\tilde{A}$ are those of A, etc.

Approximate eigenvectors: Let $\tilde{A}^{\frac{1}{2}}$ be the square root of $\tilde{A}$, and $S = \tilde{A} + \tilde{A}^{-\frac{1}{2}} \tilde{B} \tilde{B}^T \tilde{A}^{-\frac{1}{2}}$. Diagonalize S as $S = U_s \Lambda_s U_s^T$. Then $\tilde{K}$ is diagonalized by

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

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 V and projecting $\tilde{K}$ onto the normalized eigenvectors. This gives a dimensionality reduction of our images that makes the discrimination task easier.

Quality of approximation: It is difficult to verify that the approximation is accurate directly, because we are unable to form K, let alone evaluate its eigenvectors. However, we have some evidence the approximation is sound. First, the eigenvalues of $\tilde{A}$ tend to fall off quickly, despite the fact that the elements of $\mathcal{B}$ are chosen at random. This suggests that K does, indeed, have low rank. Second, in practice the representation is quite effective.

3.3.2 LDA

The i'th face image is now represented by its vector of kernel principal components $\mathbf{v}_i$. Assume that the identity of each face is known. Then we can compute linear discriminants for these vectors in the usual way [30], writing m for the number of classes, $\mathcal{C}_l$ for the set of elements in class l, N_l for the number of samples in class l, μ_l for the mean of class l, and computing the within class variance W and between class variance B 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 $\mathbf{w}$ that maximizes the ratio,

$$\mathbf{w}_{opt} = argmax_{\mathbf{w}} \frac{\mathbf{w}^T B \mathbf{w}}{\mathbf{w}^T W \mathbf{w}},$$

by solving the generalized eigenvalue problem:

$$B\mathbf{w} = \lambda W \mathbf{w}.$$

We obtain a set of projection directions, which we stack into a matrix

$$W = \begin{bmatrix} \mathbf{w}_1^T \\ \dots \\ \mathbf{w}_n^T \end{bmatrix}.$$

The final representation for the i'th face is now $\mathbf{f}_i = W\mathbf{v}_i$. Notice that, in this coordinate system, the Mahalanobis distance to a class mean is given by the Euclidean distance.

Of course, not all of our images are labeled. However, we do have a subset of data where there was a single face detected in an image with a single name in the caption. We use these images to compute linear discriminants in the first instance. Later in the process, we will have labels for each item, and can re-estimate linear discriminants.

4 Name Assignment by Simple Clustering

We have a set of b "bags", each containing F faces and N names. We wish to identify correspondences between names and faces within each bag. Each name in a bag can belong to at most one face. If we had a cluster of face vectors for each individual, we could allocate the name whose cluster center is closest to each face (this would also require allocating each name only once, and not naming a face if all cluster centers are too far away). With the allocations, we could re-estimate cluster centers, and so on. This method is analogous to k-means clustering (see the textbook account in [21], for example). There are advantages to generalizing the method with a probabilistic model: we can perform soft allocations of names to faces; we will be able to benefit from text features (section 5); and it is easier to reason explicitly about both faces without names and exclusion between names. To build a probabilistic model, we regard correspondence as a hidden variable, build a generative model for a bag given correspondence, obtain a complete data log-likelihood, and then estimate with the expectation-maximization (EM) algorithm. A variant estimation procedure, where one chooses the best correspondence rather than a weighted average of correspondences, performs better in practice.

4.1 A Generative Model for Bags

To obtain a bag of data, we first draw the number of faces F from a distribution $P(F)$ and the number of names N from a distribution $P(N)$. We then generate N names $\mathbf{n}_i$, each with a context $\mathbf{c}_i$, as IID samples of $P(\mathbf{n}, \mathbf{c})$. The context is of no interest at this point, but we will use the idea below. In turn, each name and its context generates a binary variable $pictured$, which determines whether the name will generate a face in the image. For each name for which $pictured = 1$ (the total number of such names cannot exceed F), a face $\mathbf{f}_i$ is generated from the conditional density $P(\mathbf{f}|\mathbf{n}_i, \theta)$, where θ are parameters of this distribution which will need to be estimated. The remaining faces are generated as IID samples from a distribution $P(f)$. We cannot observe which name generated which face, and must encode this information with a hidden variable.

For the moment, assume that we know a correspondence from names to faces for a particular bag. This is encoded as a partition of the names $\mathcal{N}$ in the bag into two sets, $\mathcal{D}$ being the names that generate faces and $\mathcal{U}$ being the names that do not, and a map σ, which takes a name index α to a face index $\sigma(\alpha)$. For convenience, we write the set of faces in the bag as $\mathcal{F}$. The likelihood of the bag is then

$$L(\theta, \sigma) = P(N)P(F) \left(\prod_{\alpha \in \mathcal{D}} P(\mathbf{f}_{\sigma(\alpha)}|\mathbf{n}_\alpha, \theta) \right) \left(\prod_{\gamma \in \mathcal{F} - \sigma(\mathcal{D})} P(\mathbf{f}_\gamma) \right) \left(\prod_{u \in \mathcal{N}} P(\mathbf{n}_u, \mathbf{c}_u) \right).$$

Notice that $pictured$ does not appear explicitly here (it is implicit in the form of the likelihood).

Implementation details: F and N typically vary between one and five, and we see no advantage in regarding larger bags as different from smaller ones. We therefore regard $P(N)$ and $P(F)$ as uniform over the range of those variables, and so they play no part in the estimation. We use a uniform prior over names and contexts ($P(n_u, c_u)$), too, and they too play no further part in the estimation. We regard $P(\mathbf{f}_\gamma)$ as uniform; we will use only its logarithm, which will be a constant parameter. Our choice of coordinate system means we can regard $P(\mathbf{f}|\mathbf{n}, \theta)$ as a normal distribution, with mean $\theta_{\mathbf{n}}$ — which gives one cluster center per name — and covariance $\sigma_f^2 \mathcal{I}$. We choose a sigma to produce reasonable values ($\sigma = 0.1$), but do not fit this explicitly.

4.2 Estimation with EM

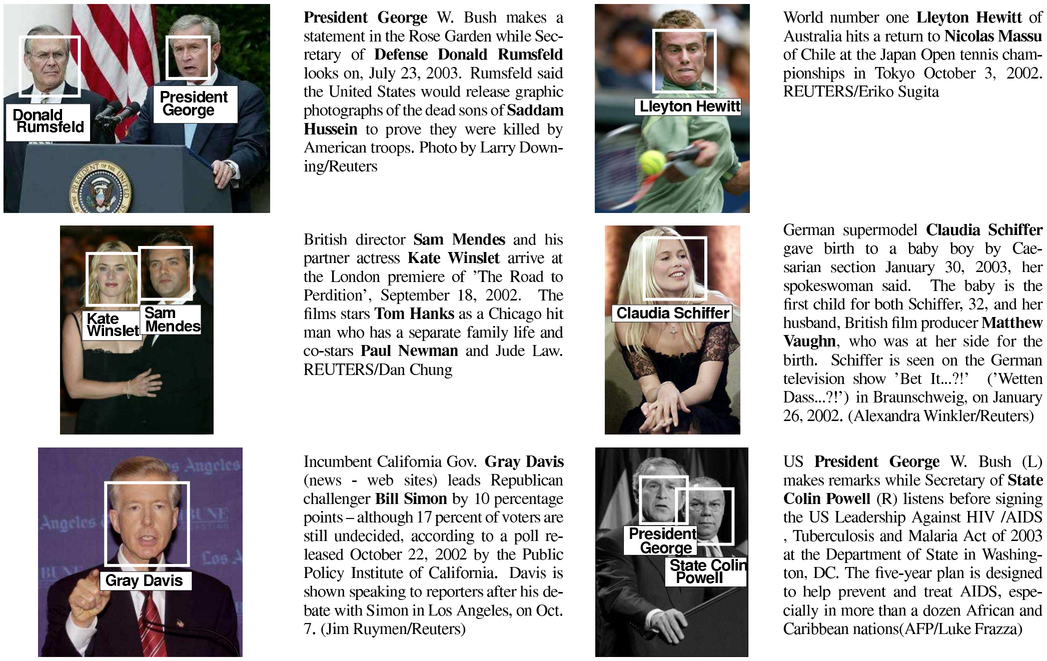
Of course, the correspondence between faces and names is unknown. However, for each bag there is a small set of possible correspondences. We construct an indicator variable $\delta(m, n)$, where

$$\delta(m, n) = \left\{ \begin{array}{cc} 1 & \text{if the } n\text{'th correspondence for the } m\text{'th data item actually occurs} \\ 0 & \text{otherwise} \end{array} \right\}$$

This indicator variable is unknown, but we will estimate it. If it were known, we could write the log-likelihood of the data set as

$$\sum_{m \in \text{data}} \left(\sum_{n \in \text{correspondences for the } m\text{'th data item}} \delta(m, n) \log L(\theta, \sigma_n) \right).$$

We now estimate θ and $\delta(m, n)$ with EM. It is natural to regard this as a soft-count procedure. At the i'th iteration, we first estimate the expected value of the $\delta(m, n)$ conditioned on the previous parameter estimate $\theta^{(i)}$, then estimate $\theta^{(i+1)}$ by substituting these expected values in equation 4.2 and maximizing. As this is a straightforward case, we omit detailed calculations.



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)

Figure 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*).*

4.3 Estimation with Maximal Assignment

If the model is an accurate reflection of the data, then it is natural to average out hidden variables (rather than, say, simply maximizing over them), and doing so should give better estimates (e.g. [44]). However, the procedure is regularly outperformed in vision problems by the simpler — and statistically non-optimal — procedure of maximizing over the hidden variables (for example, randomized search for correspondences in fundamental matrix estimation [67, 66]). We conjecture that this is because local models — in our case, $p(\mathbf{f}|\mathbf{n}, \theta)$ — may exaggerate the probability of large errors, and so the expectation step could weight poor correspondences too heavily.

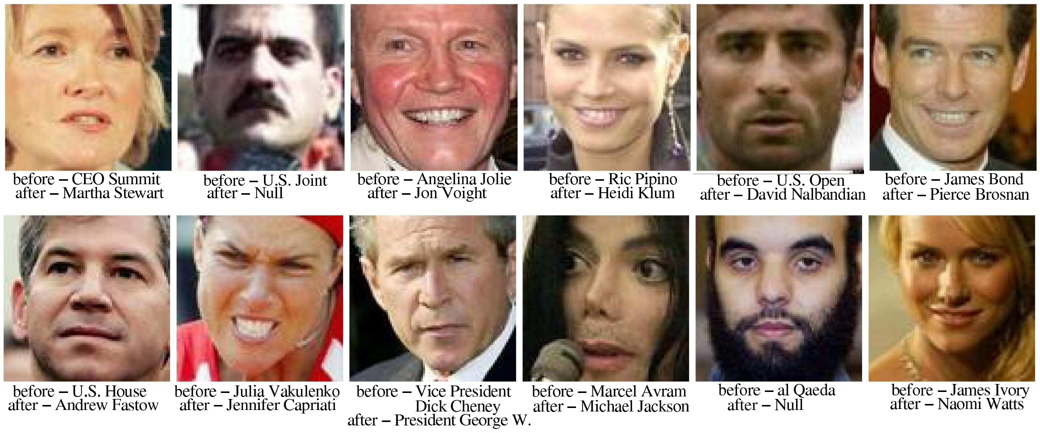
Maximal assignment iterates two steps:

- Set the $\delta(m, n)$ corresponding to the maximum likelihood assignment to 1 and all others to 0.
- Maximize the parameters $P(f|n, \theta_f)$ using counts.

In practice, maximal assignment leads to better name predictions (section 6).

5 Clustering with Context Understanding

Up to this point, we've treated the caption as a bag of words. However, the context of the caption is important. For example, consider the caption:



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

Figure 6: *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.*

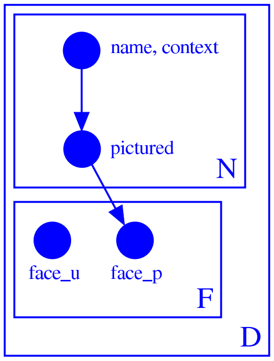
> *Sahar Aziz, left, a law student at the University of Texas, hands the business card identifying Department of the Army special agent Jason D. Treesh to one of her attorneys, Bill Allison, right, during a news conference on Friday, Feb. 13, 2004, in Austin, Texas. . . . In the background is Jim Harrington, director of the Texas Civil Rights Project. (AP Photo/Harry Cabluck)*

From the caption alone, we expect to see Sahar Aziz, Bill Allison and Jim Harrington in the picture, and we do not expect to see Jason D. Treesh. This suggests that a language model can exclude some names from consideration. In this section, we show how to build such a model into our framework (section 5.1); describe two plausible such models (section 5.2); and describe two estimation methods (sections 5.3 and 5.4).

5.1 A Generative Model for Bags

Many of the caption phenomena that suggest a person is present are relatively simple, and a simple language model should exclude some names from consideration. There are three important cases. 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. Second, caption writers tend to name people who are actually depicted earlier in the caption. Third, caption writers regularly use depiction indicators such as “left”, “(R)”, “background”.

Our generative model can be enhanced in a relatively straightforward way to take advantage of these phenomena. In section 4, we encoded $pictured$ implicitly in the correspondence. We must now recognize $pictured$ as a random variable, and incorporate it into the model. Doing so yields the following generative model:



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 $P(pictured|\text{context}, \theta_c)$).
4. For each *name, context* pair where $pictured = 1$, generate a face from $P(\mathbf{f}|\mathbf{n}, \theta_f)$.
5. Generate $F - \sum pictured$ other faces from $P(\mathbf{f})$.

We follow section 4.1 to obtain an expression for the likelihood of a bag conditioned on known correspondence. To obtain a bag of data, we first draw the number of faces F from a distribution $P(F)$ and the number of names N from a distribution $P(N)$. We then generate N names $\mathbf{n}_i$, each with a context $\mathbf{c}_i$, as IID samples of $P(\mathbf{n}, \mathbf{c})$. In turn, each name and its context generates a binary variable $pictured$, which determines whether the name will generate a face in the image, from $P(pictured|\text{context}, \theta_c)$. For each name for which $pictured = 1$ (the total number of such names cannot exceed F), a face $\mathbf{f}_i$ is generated from the conditional density $P(\mathbf{f}|\mathbf{n}_i, \theta)$, where θ are parameters of this distribution which will need to be estimated. The remaining faces are generated as IID samples from a distribution $P(f)$. We cannot observe which name generated which face, and must encode this information with a hidden variable.

For the moment, assume that we know a correspondence from names to faces for a particular bag. Notice that this implicitly encodes $pictured$: names that have corresponding faces have $pictured = 1$, and the others have $pictured = 0$. The correspondence is encoded as a partition of the names $\mathcal{N}$ in the bag into two sets, $\mathcal{D}$ being the names that generate faces and $\mathcal{U}$ being the names that do not, and a map σ, which takes a name index $\alpha \in$ to a face index $\sigma(\alpha)$. For convenience, we write the set of faces in the bag as $\mathcal{F}$. The likelihood of the bag is then

$$\begin{aligned} L(\theta_c, \theta_f, \sigma) \quad = \quad & P(N)P(F)\left(\prod_{\alpha \in \mathcal{D}} P(\mathbf{f}_{\sigma(\alpha)}|\mathbf{n}_\alpha, \theta_f)P(pictured = 1|\mathbf{c}_\alpha, \theta_c)\right) * \\ & \left(\prod_{\gamma \in \mathcal{F} - \sigma(\mathcal{D})} P(\mathbf{f}_\gamma)(P(pictured = 0|\mathbf{c}_\gamma, \theta_c))\right)\left(\prod_{u \in \mathcal{N}} P(\mathbf{n}_u, \mathbf{c}_u)\right). \end{aligned}$$

We need a model of $P(pictured = 1|\text{context}, \theta_c)$. Once we have a model, we must estimate θ_f (the parameters of the distribution generating faces from names) and θ_c (the parameters of the distribution generating $pictured$ from context). All parameters are treated as before (section 4.1), except now we also fit a model of name context, $P(pictured = 1|\mathbf{c}_\gamma, \theta_c)$.

5.2 Language Representation

We have explored two models for $P(pictured|\text{context}, \theta_c)$. First, a naive Bayes model in which each of the different context cues is assumed independent given the variable pictured, and second, 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

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

Line 1 is Bayes Rule. Line 2 follows from the naive Bayes assumption. Line 3 follows again by Bayes Rule. The Z in line 4 is dependent only on the cues $C_1, ..., C_n$. 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.

Implementation details: The cues we use are: 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.

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).

5.2.2 Maximum Entropy Model

Maximum entropy models have been used extensively in natural language systems (e.g. [12]). Maximum likelihood applied to these models — otherwise known as conditional exponential models — results in a model that is consistent with a chosen set of observed statistics of the data, but which otherwise maximizes entropy. An 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. They also do not assume independence of features as the Naive Bayes model does and model conditional distributions directly rather than through the use of Bayes’ rule.

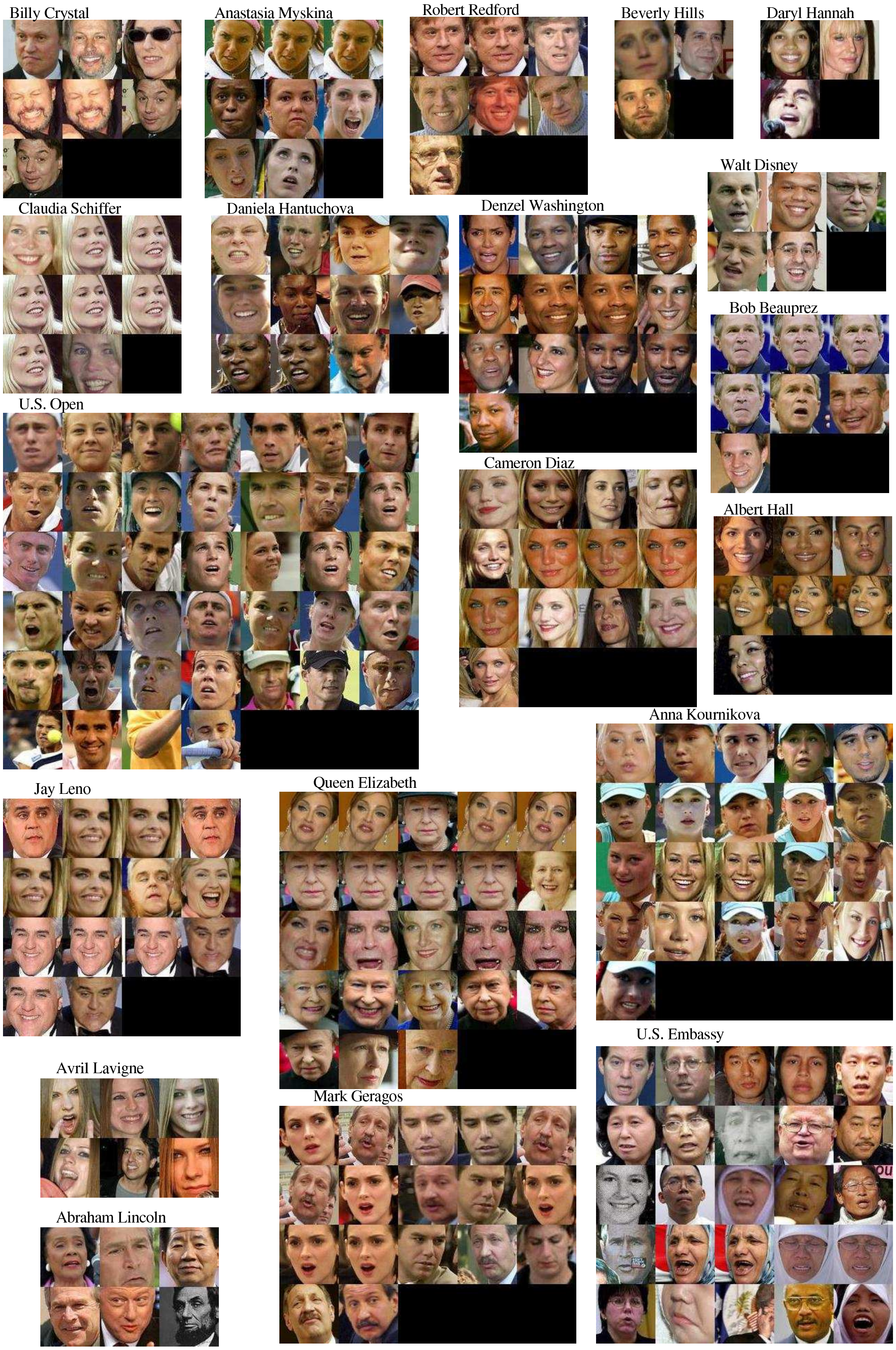

Billy Crystal
Anastasia Myskina
Robert Redford
Beverly Hills
Daryl Hannah
Walt Disney
Claudia Schiffer
Daniela Hantuchova
Denzel Washington
Bob Beauprez
U.S. Open
Cameron Diaz
Albert Hall
Anna Kournikova
Jay Leno
Queen Elizabeth
U.S. Embassy
Avril Lavigne
Mark Geragos
Abraham Lincoln

Figure 7: *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").*

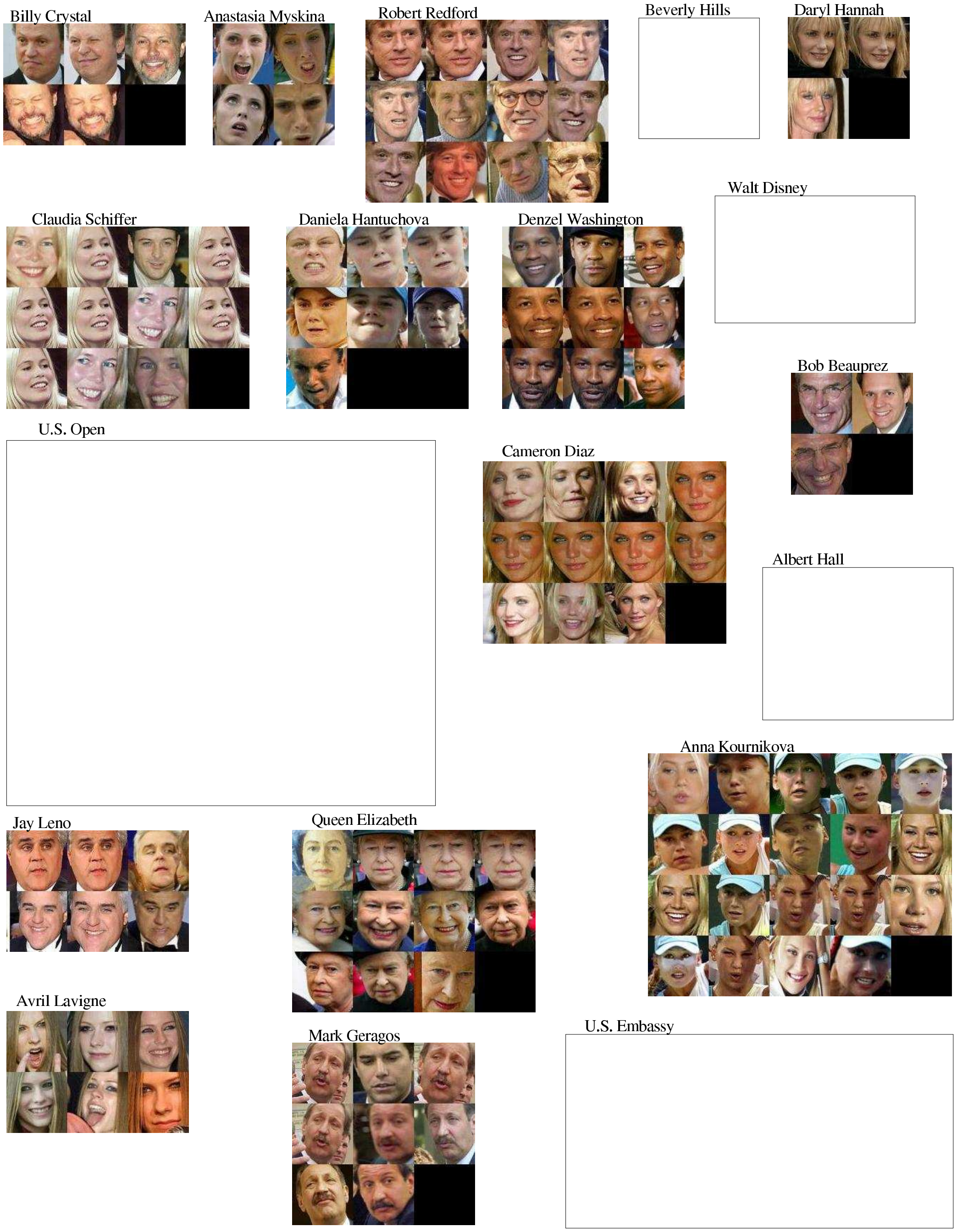

Billy Crystal
Anastasia Myskina
Robert Redford
Beverly Hills
Daryl Hannah
Walt Disney
Claudia Schiffer
Daniela Hantuchova
Denzel Washington
Bob Beauprez
U.S. Open
Cameron Diaz
Albert Hall
Anna Kournikova
Jay Leno
Queen Elizabeth
Avril Lavigne
Mark Geragos
U.S. Embassy

Figure 8: *The clusters of Figure 7 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").*

Recall $pictured$ is a binary variable. We are modeling $P(pictured = 1|\text{context}, \theta_c)$. We encode context as a binary vector, where an element of the vector is 1 if the corresponding context cue is true and zero if it is not. For the i'th context cue we define two 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 model is now

$$p(pictured|x, \theta_c) \propto exp \sum_j \theta_{c,j} f_j(x, pictured)$$

where $\theta_{c,j}$ is the weight of indicator function j.

Implementation details: 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. We also define a separate cue for each binned location corresponding to the binned location cue used for the Naive Bayes model. For the Maximum Entropy model we also add cues looking for specific strings ("pictured", "shown", "depicted" and "photo").

5.3 Estimation with EM

EM is computed as described in section 4.2. The differences for each context model are described in section 5.3.1 and section 5.4.

5.3.1 Estimating Depiction with Naive Bayes

We update the distributions, $P(pictured|C_i)$ and $P(pictured)$, at each iteration of EM process using maximum likelihood estimates based on soft counts. $P(pictured|C_i)$ is updated by 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.

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.

5.3.2 Estimating Depiction with Maximum Entropy Models

To find the maximum likelihood $p(y|x)$, we use improved iterative scaling, the standard algorithm for finding maximum entropy distributions, again using soft counts. Details of this model and algorithm are described in [12].

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

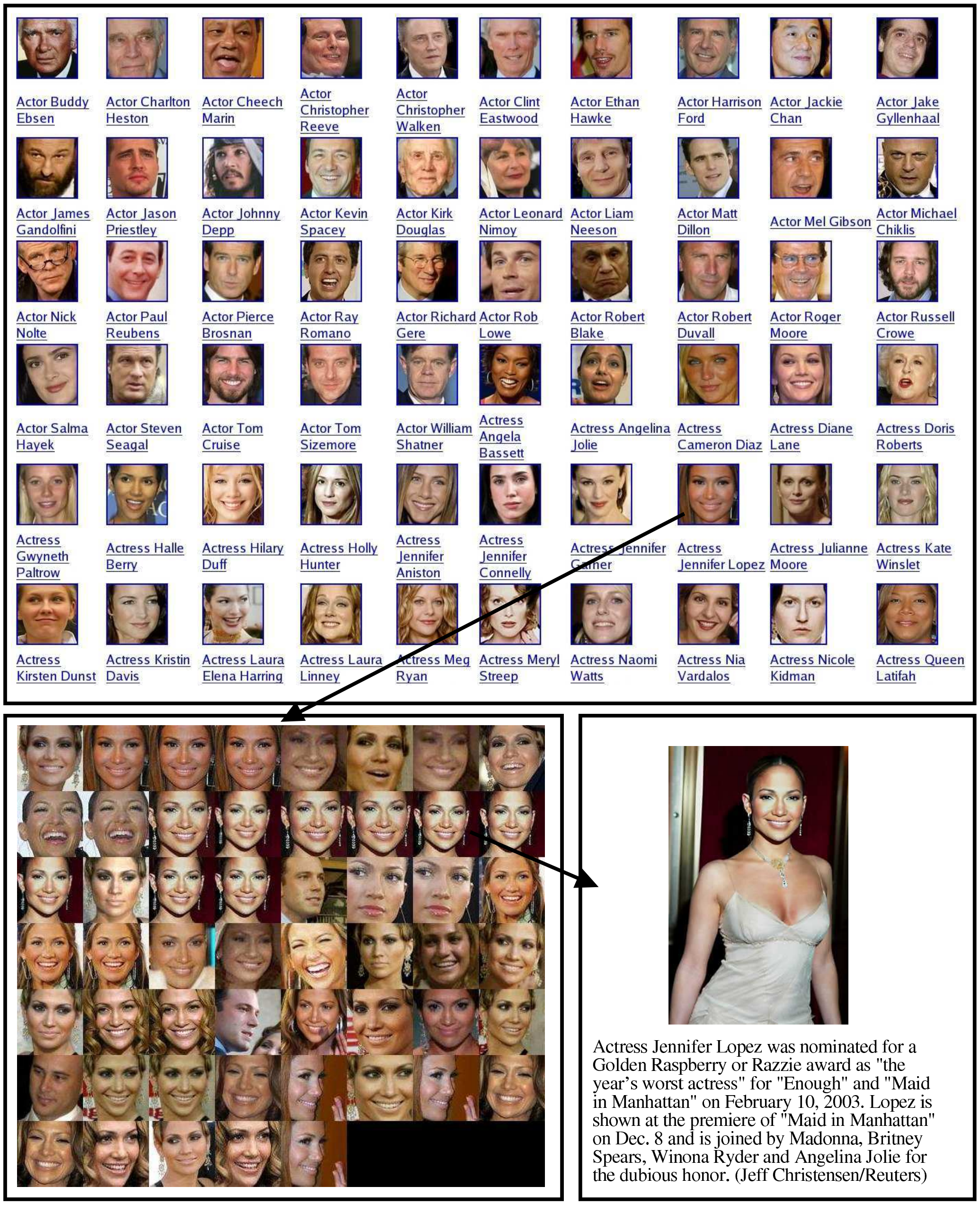
Figure 9: *Our new procedure gives us not only better clustering results, but also a natural language classifier which can be tested separately.* **Above:** *a few captions where detected names have been labeled with IN (pictured) and OUT (not pictured) using our learned language model. Our language 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".*

5.4 Estimation with Maximal Assignment

Estimation with maximal assignment is as before. However, both naive Bayes and maximum entropy language models no longer use soft counts. In effect, maximal assignment chooses a single correspondence, and so specifies which names are depicted. The conditional language models and appearance models are then learned with supervised data (it is known for every context whether it is depicted or not and also which face has been assigned to each name) using maximum likelihood.

6 Results

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 data set. 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 compel our method to associate a single name with each face (we use the name given by the maximum likelihood correspondence once the parameters have been estimated), and then determine how many faces in the evaluation set are correctly labeled by


Actor Buddy Ebsen
Actor Charlton Heston
Actor Cheech Marin
Actor Christopher Reeve
Actor Christopher Walken
Actor Clint Eastwood
Actor Ethan Hawke
Actor Harrison Ford
Actor Jackie Chan
Actor Jake Gyllenhaal
Actor James Gandolfini
Actor Jason Priestley
Actor Johnny Depp
Actor Kevin Spacey
Actor Kirk Douglas
Actor Leonard Nimoy
Actor Liam Neeson
Actor Matt Dillon
Actor Mel Gibson
Actor Michael Chiklis
Actor Nick Nolte
Actor Paul Reubens
Actor Pierce Brosnan
Actor Ray Romano
Actor Richard Gere
Actor Rob Lowe
Actor Robert Blake
Actor Robert Duvall
Actor Roger Moore
Actor Russell Crowe
Actor Salma Hayek
Actor Steven Seagal
Actor Tom Cruise
Actor Tom Sizemore
Actor William Shatner
Actress Angela Bassett
Actress Angelina Jolie
Actress Cameron Diaz
Actress Diane Lane
Actress Doris Roberts
Actress Gwyneth Paltrow
Actress Halle Berry
Actress Hilary Duff
Actress Holly Hunter
Actress Jennifer Aniston
Actress Jennifer Connelly
Actress Jennifer Garner
Actress Jennifer Lopez
Actress Julianne Moore
Actress Kate Winslet
Actress Kirsten Dunst
Actress Kristin Davis
Actress Laura Elena Harring
Actress Laura Linney
Actress Meg Ryan
Actress Meryl Streep
Actress Naomi Watts
Actress Nia Vardalos
Actress Nicole Kidman
Actress Queen Latifah

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)

Figure 10: *We have created a web interface for organizing and browsing news photographs according to individual. Our data set 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.*

Model	EM	MM
Baseline PCA Appearance Model, No Lang Model	$37 \pm .04\%$	$53 \pm .04\%$
kPCA+LDA Appearance Model, No Lang Model	$56 \pm .05\%$	$67 \pm .03\%$
kPCA+LDA Appearance Model + N.B. Lang Model	$72 \pm .04\%$	$77 \pm .04\%$
kPCA+LDA Appearance Model + Max Ent Lang Model	–	$78 \pm .04\%$

Table 1: **Above:** *To form an evaluation set, we randomly selected 1000 faces from our data set 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) as well as for a baseline PCA appearance model. Standard deviation is calculated by dividing the test set into 10 subsets containing 100 faces each and calculating the deviation over the accuracies for these subsets. 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.*

that name. This is a stern test; a less demanding alternative is to predict a ranked list of names for a given face, but this is harder to evaluate.

kPCA+LDA is a reasonable model: We test our appearance model against a commonly used baseline face representation of principal components analysis. In table 1 we see that the appearance only clustering using kPCA followed by LDA performs better than the PCA appearance model. kPCA plus LDA labels 67% of the faces correctly, while PCA labels 53% of the faces correctly.

Maximal assignment performs better than EM: In table 1, we see that the basic clustering correctly labels 56% of the test images correctly when estimated with EM (as in section 4.2), and 67% of the test images correctly when estimated with maximal assignment (as in section 4.3). For context understanding clustering, 72% of the faces are labeled correctly when estimated with EM (section 5.3), where as 77% of the faces are labeled correctly when estimated with maximal assignment (section 5.4). This clearly indicates that the maximal assignment procedure performs better than EM for our labeling task. We speculate that the Gaussian model of face features conditioned on a name places too much weight on faces that are far from the mean. One other possible explanation for this phenomenon is that MM is training the model under the exact conditions for which it is tested on (to get the top correspondence correct). It would be interesting to measure the average log probability of the correct correspondence on the evaluation set, which is what EM optimizes.

Language cues are helpful: Language cues are helpful, because they can rule out some bad labelings. Using the same test set, we see that context understanding clustering (section 5) labels 77% of the test faces correctly using a naive Bayes model and 78% of the faces correctly using a maximum entropy model (table 1).

Vision reinforces language: One consequence of our context understanding clustering method is a pure natural language understanding module, which can tell whether faces are depicted in captions from context alone (i.e. one looks at $P(pictured = 1|\mathbf{c})$). We expect that, if context understanding clustering works, this module should be reasonably accurate. The module is, indeed, accurate. 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

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 data set 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.*

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 data set. Figure 9 shows some example captions labeled using the learned Maximum Entropy Context model. 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.

Spatial context: One could reasonably expect that caption features like "(left)" might directly suggest a correspondence, rather than just indicate depiction. However, incorporating this information into our context understanding model was not particularly helpful. In particular, we we built 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.

Scale: The most natural comparison with our work is that of Yang *et al.* ([76], and described briefly above). This work applies various multiple-instance learning methods to learn the correct association of name to face for bags consisting of a single face and 4.7 names on average. There are 234 bags where the correct name appears in the bag, and 242 where it does not; methods label between 44% and 63% of test images correctly, depending on the method. Our method shows appreciable improvements. We conjecture that there are two places in which operating with very large scale data sets is helpful. First, kPCA estimates seem to give better representations when more images are used, perhaps because high-variance directions are more stably identified. Second, more data appears to simplify correspondence

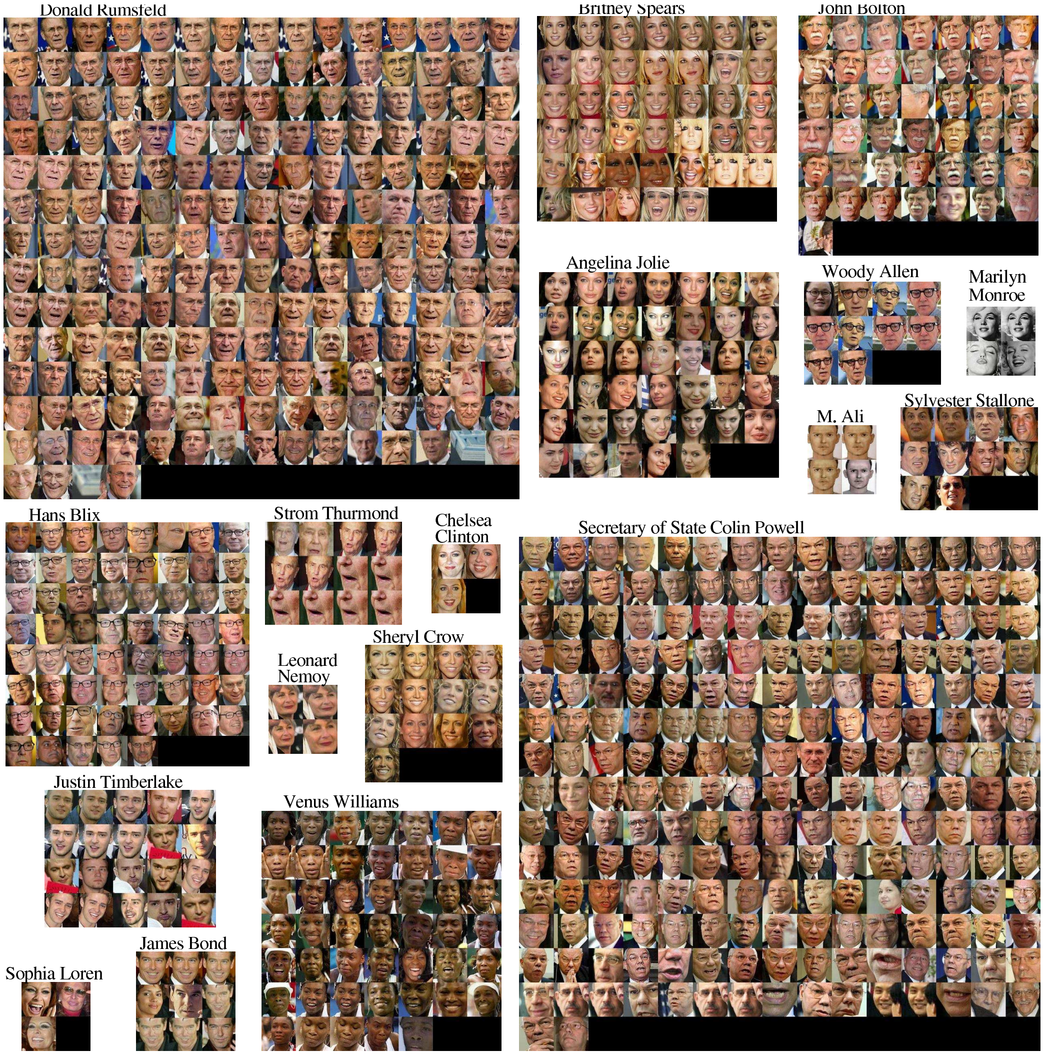

Donald Rumsfeld
Britney Spears
John Bolton
Angelina Jolie
Woody Allen
Marilyn Monroe
M. Ali
Sylvester Stallone
Hans Blix
Strom Thurmond
Chelsea Clinton
Secretary of State Colin Powell
Leonard Nemoy
Sheryl Crow
Justin Timberlake
Venus Williams
James Bond
Sophia Loren

Figure 11: *The figure shows a representative set of clusters, illustrating a series of important properties of both the data set 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*).*

problems, because the pool of relatively easily labeled images will grow. Such images might consist of faces that happen to have only one possible label, or of groups of faces where there is little doubt about the labeling (for example, two faces which are very different; as another example, one familiar face and its name together with an unfamiliar face and its name). We conjecture that the absolute size of the "easy" set is an important parameter, because a large set of easy images will make other images easy to label. For example, an image that contains two unfamiliar faces and two unfamiliar names could be much easier to label if, in another image, one of these faces appeared with a familiar face. If this conjecture is true, the problem simplifies as one operates with larger data sets.

6.1 Recognition Baselines

We have performed several baseline recognition tests to measure the difficulty of the face recognition data set produced by our system. To do this, we select 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 data sets 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 data sets, suggesting that our face images are in fact quite challenging and that they will be a useful data set for training future face recognition systems.

7 Conclusion

We have automatically produced a very large and realistic face data set consisting of 30,281 faces with roughly 3,000 different individuals from news photographs with associated captions. This data set 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 data set. 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 5.4), 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] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Proc. NIPS 15*, pages 561–568. MIT Press, 2003.

[2] O. Arandjelovic and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.

[3] K. Barnard, P. Duygulu, and D.A. Forsyth. Clustering art. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages II:434–441, 2001.

[4] K. Barnard and D.A. Forsyth. Learning the semantics of words and pictures. In *Int. Conf. on Computer Vision*, pages 408–15, 2001.

[5] Kobus Barnard, Pinar Duygulu, Nando de Freitas, David Forsyth, David Blei, and Michael I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[6] Kobus Barnard, Pinar Duygulu, Raghavendra Guru, Prasad Gabbur, , and David Forsyth. The effects of segmentation and feature choice in a translation model of object recognition. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2003.

[7] Kobus Barnard and Matthew Johnson. Word sense disambiguation with pictures. *Artif. Intell.*, 167(1-2):13–30, 2005.

[8] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class-specific linear projection. *IEEE T. Pattern Analysis and Machine Intelligence*, 19(7):711–720, July 1997.

[9] A.C. Berg, T.L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *CVPR05*, pages I:26–33, 2005.

[10] A.C. Berg and J. Malik. Geometric blur and template matching. In *CVPR01*, pages I:607–614, 2001.

[11] T.L. Berg and D.A. Forsyth. Animals on the web. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 1463–1470, 2006.

[12] A. Berger, S.D. Pietra, and V. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1), 1996.

[13] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE T. Pattern Analysis and Machine Intelligence*, 25(9), 2003.

[14] David M. Blei and Michael I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, New York, NY, USA, 2003. ACM Press.

[15] P. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 32(2):263–311, 1993.

[16] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld – image segmentation using expectationmaximization and its application to image querying. *IEEE T. Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, 2002.

[17] Yixin Chen and James Z. Wang. Image categorization by learning and reasoning with regions. *J. Mach. Learn. Res.*, 5:913–939, 2004.

[18] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. Gate: A framework and graphical development environment for robust nlp tools and applications. In *40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[19] Ritendra Datta, Jia Li, and James Z. Wang. Content-based image retrieval: approaches and trends of the new age. In *MIR '05: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 253–262, New York, NY, USA, 2005. ACM Press.

[20] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1-2):31–71, 1997.

[21] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. 2001.

[22] P. Duygulu, K. Barnard, N. de Freitas, and D.A. Forsyth. Object recognition as machine translation. In *Proc. European Conference on Computer Vision*, pages IV: 97–112, 2002.

[23] Andras Ferencz, Erik Learned-Miller, and Jitendra Malik. Learning hyper-features for visual identification. In *Advances in Neural Information Processing*, volume 18, 2005.

[24] A.W. Fitzgibbon and A. Zisserman. On affine invariant clustering and automatic cast listing in movies. In *Proc. 7th European Conference on Computer Vision*. Springer-Verlag, 2002.

[25] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nyström method. *IEEE T. Pattern Analysis and Machine Intelligence*, 25(2), 2004.

[26] G.H. Golub and C.F. Van Loan. *Matrix Computations*. The Johns Hopkins Unversity Press, Baltimore, third edition, 1996.

[27] V. Govindaraju, D.B. Sher, R.K. Srihari, and S.N. Srihari. Locating human faces in newspaper photographs. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 549–554, 1989.

[28] R. Gross, I. Matthews, and S. Baker. Appearance-based face recognition and light-fields. *IEEE T. Pattern Analysis and Machine Intelligence*, 26(4):449– 465, 2004.

[29] R. Gross, J. Shi, and J. Cohn. Quo vadis face recognition? In *Third Workshop on Empirical Evaluation Methods in Computer Vision*, 2001.

[30] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Verlag, 2001.

[31] Thomas Hofmann and Jan Puzicha. Statistical models for co-occurrence data. A.I. Memo 1635, Massachusetts Institute of Technology, 1998.

[32] Ricky Houghton. Named faces: Putting names to faces. *IEEE Intelligent Systems*, 14(5):45–50, 1999.

[33] S. Ioffe and D. Forsyth. Mixtures of trees for object recognition. In *CVPR*, 2001.

[34] V. Jain, A. Ferencz, and E. Learned-Miller. Discriminative training of hyper-feature models for object identification. In *British Machine Vision Conference*, pages 357–366, 2006.

[35] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR*, pages 119–126, 2003.

[36] Dhiraj Joshi, James Z. Wang, and Jia Li. The story picturing engine: finding elite images to illustrate a story using mutual reinforcement. In *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 119–126, New York, NY, USA, 2004. ACM Press.

[37] K. I. Kim, K. Jung, and H. J. Kim. Face recognition using kernel principal component analysis. *Signal Processing Letters*, 9(2):40–42, 2002.

[38] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Neural Information Processing Systems*, 2003.

[39] Jia Li and James Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(10), 2003.

[40] J.S. Liu and R. Chen. Sequential monte-carlo methods for dynamic systems. Technical report, Stanford University, 1999. preprint.

[41] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos. Face recognition using kernel direct discriminant analysis algorithms. *IEEE Trans. Neural Networks*, 14(1):117– 126, 2003.

[42] O. Maron and A.L. Ratan. Multiple-instance learning for natural scene classification. In *The Fifteenth International Conference on Machine Learning*, 1998.

[43] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 570–576, Cambridge, MA, USA, 1998. MIT Press.

[44] G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. John Wiley and Sons, 1996.

[45] K. Mikolajczyk. Face detector. Technical report, INRIA Rhone-Alpes. Ph.D report.

[46] Mor Naaman, Ron B. Yeh, Hector Garcia-Molina, and Andreas Paepcke. Leveraging context to resolve identity in photo albums. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 178–187, New York, NY, USA, 2005. ACM Press.

[47] D. Ozkan and P. Duygulu. A graph based approach for naming faces in news photos. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages II: 1477–1482, 2006.

[48] P. J. Phillips and E. Newton. Meta-analysis of face recognition algorithms. In *Proceeedings of the Int. Conf. on Automatic Face and Gesture Recognition*, 2002.

[49] P.J. Phillips, P. Grother, R.J Micheals, D.M. Blackburn, E Tabassi, and J.M. Bone. Frvt 2002: Evaluation report. Technical report, Face Recognition Vendor Test, 2002.

[50] T. Poggio and Kah-Kay Sung. Finding human faces with a gaussian mixture distribution-based face model. In *Asian Conf. on Computer Vision*, pages 435–440, 1995.

[51] Soumya Ray and Mark Craven. Supervised versus multiple instance learning: an empirical comparison. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 697–704, New York, NY, USA, 2005. ACM Press.

[52] H.A. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing 8*, pages 875–881, 1996.

[53] H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 203–8, 1996.

[54] H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE T. Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[55] H.A. Rowley, S. Baluja, and T. Kanade. Rotation invariant neural network-based face detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 38–44, 1998.

[56] S. Satoh, Y. Nakamura, and T. Kanade. Name-it: naming and detecting faces in news videos. *IEEE Multimedia*, 6(1):22–35, 1999.

[57] Shin'ichi Satoh and Takeo Kanade. Name-it: Association of face and name in video. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 368, Washington, DC, USA, 1997. IEEE Computer Society.

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

[59] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *CVPR00*, pages I: 746–751, 2000.

[60] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[61] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of Optical Society of America*, 4(3):519–524, 1987.

[62] Xiaodan Song, Ching-Yung Lin, and Ming-Ting Sun. Cross-modality automatic face model training from large video databases. In *CVPRW '04: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 5*, page 91, Washington, DC, USA, 2004. IEEE Computer Society.

[63] R.K. Srihari. Automatic indexing and content based retrieval of captioned images. *Computer*, 28(9):49–56, 1995.

[64] K-K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE T. Pattern Analysis and Machine Intelligence*, 20:39–51, 1998.

[65] Qingping Tao, Stephen Scott, N. V. Vinodchandran, and Thomas Takeo Osugi. Svm-based generalized multiple-instance learning via approximate box counting. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 101, New York, NY, USA, 2004. ACM Press.

[66] P. Torr and A. Zisserman. Robust computation and parametrization of multiple view relations. In *Int. Conf. on Computer Vision*, pages 485–491, 1998.

[67] P.H.S. Torr and D.W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *Int. J. Computer Vision*, 24:271–300, 1997.

[68] M. Turk and A. Pentland. Eigen faces for recognition. *J. of Cognitive Neuroscience*, 1991.

[69] P. Viola and M.J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, May 2004.

[70] H. Wactlar, T. Kanade, M. Smith, and S. Stevens. Intelligent access to digital video: The informedia project. *IEEE Computer*, 29(5), 1996.

[71] Y. Wang, Z. Liu, and J.-C. Huang. Multimedia content analysis-using both audio and visual clues. *Signal Processing Magazine*, 17(6):12–36, 2000.

[72] C. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *Proc. NIPS*, volume 13, pages 682–688, 2001.

[73] Keiji Yanai and Kobus Barnard. Image region entropy: a measure of "visualness" of web images associated with one concept. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 419–422, New York, NY, USA, 2005. ACM Press.

[74] J. Yang, A.F. Frangi, J.-Y. Yang, D. Zhang, and Z. Jin. Kpca plus lda: a complete kernel fisher discriminant framework for feature extraction and recognition. *IEEE T. Pattern Analysis and Machine Intelligence*, 27(2):230–244, 2005.

[75] Jun Yang and Alexander G. Hauptmann. Naming every individual in news video monologues. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 580–587, New York, NY, USA, 2004. ACM Press.

[76] Jun Yang, Rong Yan, and Alexander G. Hauptmann. Multiple instance learning for labeling faces in broadcasting news video. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 31–40, New York, NY, USA, 2005. ACM Press.

[77] M.-H. Yang, N. Ahuja, and D. J. Kriegman. Face recognition using kernel eigenfaces. In *IEEE Int. Conf. Image Processing*, volume 1, pages 37–40, 2000.

[78] M.H. Yang, D.J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *PAMI*, 24(1):34–58, January 2002.

[79] Ming-Hsuan Yang. Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. In *FGR '02: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, page 215, Washington, DC, USA, 2002. IEEE Computer Society.

[80] Lei Zhang, Longbin Chen, Mingjing Li, and Hongjiang Zhang. Automated annotation of human faces in family albums. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 355–358, New York, NY, USA, 2003. ACM Press.

[81] Lei Zhang, Yuxiao Hu, Mingjing Li, Weiying Ma, and Hongjiang Zhang. Efficient propagation for face annotation in family albums. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 716–723, New York, NY, USA, 2004. ACM Press.

[82] Q. Zhang and S. Goldman. Em-dd: An improved multiple-instance learning technique. In *Proc NIPS*, pages 1073–1080, 2001.

[83] Z. Zhang, R.K. Srihari, and A. Rao. Face detection and its applications in intelligent and focused image retrieval. In *11'th IEEE Int. Conf. Tools with Artificial Intelligence*, pages 121–128, 1999.

[84] L. Zhao, K. Kipper, W. Schuler, C. Vogler, N. Badler, and M. Palmer. Machine translation system from english to american sign language. In *Association for Machine Translation in the Americas*, 2000.

[85] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.