Improving People Search Using Query Expansions: How Friends Help To Find People

Thomas Mensink, Jakob Verbeek

Improving People Search Using Query Expansions
How Friends Help To Find People

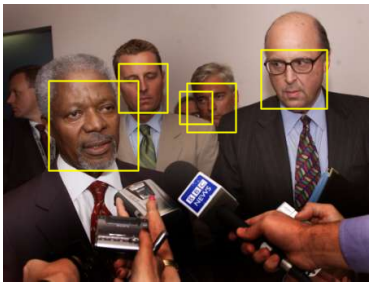
Thomas Mensink and Jakob Verbeek

LEAR - INRIA Rhône Alpes - Grenoble, France
{thomas.mensink,jakob.verbeek}@inria.fr

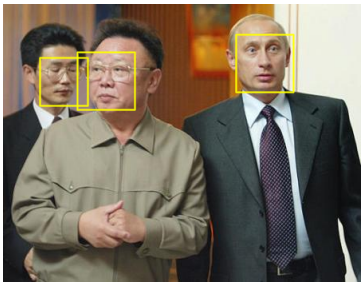
Abstract. In this paper we are interested in finding images of people on the web, and more specifically within large databases of captioned news images. It has recently been shown that visual analysis of the faces in images returned on a text-based query over captions can significantly improve search results. The underlying idea to improve the text-based results is that although this initial result is imperfect, it will render the queried person to be relatively frequent as compared to other people, so we can search for a large group of highly similar faces. The performance of such methods depends strongly on this assumption: for people whose face appears in less than about 40% of the initial text-based result, the performance may be very poor. The contribution of this paper is to improve search results by exploiting faces of other people that co-occur frequently with the queried person. We refer to this process as 'query expansion'. In the face analysis we use the query expansion to provide a query-specific relevant set of 'negative' examples which should be separated from the potentially positive examples in the text-based result set. We apply this idea to a recently-proposed method which filters the initial result set using a Gaussian mixture model, and apply the same idea using a logistic discriminant model. We experimentally evaluate the methods using a set of 23 queries on a database of 15.000 captioned news stories from *Yahoo! News*. The results show that (i) query expansion improves both methods, (ii) that our discriminative models outperform the generative ones, and (iii) our best results surpass the state-of-the-art results by 10% precision on average.

1 Introduction

Over the last decade we have witnessed an explosive growth of image and video data available both on-line and off-line. This has lead to the need for methods to index, search, and manipulate such data in a semantically meaningful manner; bridging the apparent gap between low-level features and semantics [1]. Much research has addressed this problem using so called 'supervised' techniques that require explicit manual annotations to establish correspondences between low-level features and semantics, these correspondences are then captured in models that generalize the correspondence to other images. For example, image categorization has made significant progress using this approach [2]. Learning semantic relations from weaker forms of supervision is currently an active and broad line of research. Work along these lines includes learning correspondence between keywords and image regions [3,4], and learning image retrieval and auto-annotation with keywords [5,6]. In this work images were labeled with



United Nations Secretary General ***Kofi Annan*** stands with U.N. Security Council President and U.S. Ambassador to the U.N. **John D. Negroponte** as Annan ...



North Korean leader ***Kim Jong Il***, and Russian President **Vladimir Putin** walk after talks in Vladivostok, Friday, Aug. 23, 2002. North Korean leader Kim ...

Fig. 1. Two example images with captions. The queried person is marked in italic, detected named entities in bold, and detected faces are marked by yellow rectangles.

multiple keywords per image, requiring resolution of correspondences between image content and semantic categories. Supervision from even weaker forms of annotation are also explored, e.g. based on images and accompanying text [7,8], and video with scripts and subtitles [9,10].

In this paper we aim to improve search for people in databases of captioned news photographs; see Figure 1 for two examples taken from *Yahoo! News*. Automatic analysis of news streams is important as they are major sources in the information need of people, and news articles are published at a high frequency. When searching for images of a certain person, a simple system could (i) query the database for captions containing the the name, and (ii) rank or filter the result images by the confidence level of a face detector. An example of such a system is Google Portrait [11], which queries Google with the name and then applies a fast face detector. Recently such a system has also been integrated in Google's advanced image search. Although such a system correctly rejects images without faces (or at least those without face detections), the performance of such a system is clearly limited by the fact that it returns all images with detected faces, not only those depicting the queried person.

Identification of faces in news photographs is a challenging task, significantly more so than recognition in the usual controlled setting of face recognition: we have to deal with imperfect face detection and alignment procedures, and also with great changes in pose, expression, and lighting conditions, and poor image resolution and quality. Perhaps even more importantly, labeled data sets for learning classifiers are not generally available, and tedious to produce. However, these difficulties are partly compensated by the information contained in captions. Recently it was shown that initial text-based results can be significantly improved by filtering faces on the basis of visual features [12,13]. While using the caption alone leads to a disappointing precision of 44% (fraction of faces belonging to the queried person among all returned faces, averaged over queries for 23 people), adding face analysis increases average precision to 71%, at a recall of 85%. Others [14] have considered resolving all face-name associations in

databases of captioned news images. The potential advantage of solving name-face associations for multiple names at once is that the faces associated with one name may resolve ambiguities for other names.

We explore the middle ground between solving the complete name-face association problem, and analysis of only the initial result set. We do so by starting with a set of news stories found by querying the captions with a name, which we refer to as the query set. We then extend the query set by querying the database for names that appear frequently together with the queried person; we refer to these people as 'friends' of the queried person. We use the 'query expansion' —the set of faces found in images with friends appearing in the caption— to obtain a notion of whom we are *not* looking for.

We apply this idea to a generative mixture model to filter the text-based results, as well as to a linear discriminant method to filter the initial results. We find that query expansion gives dramatic improvements in the failure mode of existing work: cases where the queried person accounts for less than 40% of the query set. We find that both the generative and discriminative method benefit from query expansion, and that the discriminative model performs best albeit being computationally more demanding.

In the next section we discuss related work and the idea of query expansion in more detail. In section Section 3 we then describe the baseline methods; their counterparts using query expansion follow in Section 4. We present our experimental results in Section 5, and our conclusions in Section 6.

2 Related Work on Finding People in News Images

Previous work on finding faces of specific people analyzed only the faces returned from a text-based query over the captions. The assumption underlying these methods is that the query set consists of a large group of highly similar faces of the queried person, plus faces of many other people appearing each just a few times. The goal is thus to find a single coherent compact cluster in a space that also contains many outliers. A graph-based method was proposed in [12]: nodes represent faces in the query set, and edges encode similarity between faces. The faces in the subset of nodes with maximum density are returned as the faces representing the queried person. The density of a subset of nodes is defined as the number of edges that do not exit the subset divided by the number of nodes in the subset. In [13] we compared this method to one based on a Gaussian mixture model. One Gaussian is fitted on all the faces in the query set, while a second Gaussian is fitted to a subset of faces that are believed to represent the queried person. To deal with the fact that it is not known which faces belonging to the queried person, the EM algorithm is used to fit the model, under the constraint that at most one face per image can depict the queried person.

We found performance of these methods to deteriorate strongly as the frequency of the queried person in the result set drops below about 40%, contradicting their underlying assumption. In this case the faces of the queried person are obscured by many faces of other people, some of which will probably also occur quite often, as we expect strong correlations on which people co-occur on news images.

Methods that aim at solving the complete name-face association problem are motivated by the idea that finding name-face associations for all names jointly can resolve

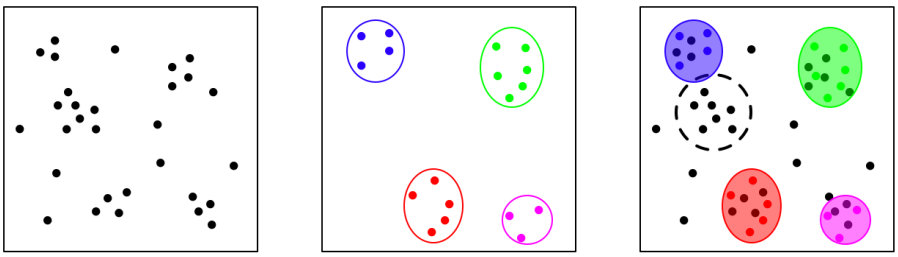


Fig. 2. Schematic illustration of how friends help to find people. The distribution of face features in the query-set (left), the query expansion with color coded faces of four people that co-occur with the queried person (middle), and how models of these people help to identify which faces in the query set are not the queried person (right).

ambiguities that would otherwise arise [14]. They work by applying a named entity detector on the captions, and a face detector on the images, and then finding the associations. The associations are constrained to assign at most one name to each face, and vice-versa. Thus if one face matches well with the model associated with a particular name, this provides evidence that the other faces should be associated with other names. In this manner documents that do not contain the name of the queried person in the caption also help to solve the query, as they provide data to learn the models for other names that co-occur with the queried person.

However, solving the complete name-face association problem is complex: captions may contain multiple names, and images multiple faces, which together have a combinatorial number of possible associations. Allowing at most one name to be associated with each face and vice-versa, a document with 6 faces and 7 names already yields 37633 possible associations. In our implementation of such a system, that relies solely on detected faces and names and does not use an additional language model, we find it to produce results that are worse than those obtained with a system that only tries to find the queried person: 54% precision with 61% recall, compared to 69% and 77% precision with 75% recall reported in [12] and [13] respectively. The worse performance of the full name-face associations is due to local maxima during learning and insufficient training data to make reliable estimates of the models for people appearing infrequently.

In this paper we aim to combine the best aspects of both approaches: (i) We *do* extend the initial query set, but only with a limited set of *relevant* stories, and model only people that co-appear relatively often. (ii) We *do not* solve the complex full name-face association problem –not even over the query expansion– but only determine in each image which (if any) face represents the queried person. By querying the database for names that frequently appear together with the queried name we collect faces that help understand whom we are *not* looking for. For example, suppose that in captions for the query *Tony Blair* the names *George Bush* and *Gordon Brown* occur often. By querying for faces of *George Bush* and *Gordon Brown* we can then rule out faces in the query set that are very similar to many faces returned for *George Bush* or *Gordon Brown*. See Figure 2 for a schematic illustration of the idea.

Our use of query expansion differs from other work using query expansion for document and image retrieval [15,16], where query expansion is used to re-query the database to obtain more similar documents or images. On the contrary, we use query expansion to obtain faces of related, but different, people to help us identify the queried person by contrasting him against these.

3 Basic Methods to Filter Text-based Query Results

Below we describe our two baseline methods to filter the query set obatined using text-based search. Both methods allow at most one face from each image to be identified as the queried person. While sometimes violated, several cases exist where a person is photographed against a background that contains the face of the person on eg. a poster, this constraint improves the results considerably.

3.1 Face Filtering using a Gaussian Mixture Model

A Gaussian mixture model has been previously used to solve the complete name-face association task by framing the problem as a constrained clustering problem [14]. Each cluster represents a name and the facial features of the person are modeled by a single Gaussian; an additional Gaussian models faces not assigned to any name. In the case of single-person querying, this approach results in a Gaussian mixture with just two mixture components [13]: one foreground model representing the queried person, and one generic face model. Below, we present the model in detail.

For each document in the query set we introduce an (unknown) assignment variable γ to represent which, if any, face in the image belongs to the queried person. Clearly, for a document with F face detections in the image, the number of assignments is just $(F+1)$: selecting one of the faces, or none ($\gamma = 0$). We then define a mixture model over the features of the detected faces $\mathcal{F} = \{f_1, \ldots, f_F\}$, marginalizing over the assignment variable γ. We use a prior over γ which is uniform over all non-zero assignments, i.e. $p(\gamma = 0) = \pi$ and $p(\gamma) = (1 - \pi)/F$ for $i \neq 0$.

$$p(\mathcal{F}) = \sum_{\gamma=0}^{F} p(\gamma)p(\mathcal{F}|\gamma), \tag{1}$$

$$p(\mathcal{F}|\gamma) = \prod_{i=1}^{F} p(f_i|\gamma), \tag{2}$$

$$p(f_i|\gamma) = \begin{cases} p_{\text{BG}}(f_i) = \mathcal{N}(f_i; \mu_{\text{BG}}, \Sigma_{\text{BG}}) & \text{if } \gamma \neq i \\ p_{\text{FG}}(f_i) = \mathcal{N}(f_i; \mu_{\text{FG}}, \Sigma_{\text{FG}}) & \text{if } \gamma = i \end{cases} \tag{3}$$

The parameters of the generic face model p_{BG} are set to the mean and variance of the faces in the query set, and we use the EM algorithm to obtain a maximum likelihood estimate the remaining parameters $\{\pi, \mu_{\text{FG}}, \Sigma_{\text{FG}}\}$. We initialize the EM algorithm in the E-step by using uniform responsibilities over the assignments: in this way faces in documents with few faces are relatively more important than faces in documents with many faces. After parameter optimization, we use the assignment maximizing $p(\gamma|\mathcal{F})$ to determine which, if any, face represents the queried person.

3.2 Face Filtering Using Linear Discriminant Analysis

Our motivation for the linear discriminant method, is to improve over the Gaussian mixture method, without resorting to a method based on pairwise similarities as in [12,13], which would be computationally costly when the query set contains many faces. We chose to use sparse multinomial logistic regression (SMLR) [17] since we are using high-dimensional face features (1664 dimensional, see Section 5). SMLR is a multi-class classifier, which is useful when considering query expansions in the next section. SMLR has been reported to perform equal or better than Support Vector Machines and Relevance Vector Machines on different benchmark classification problems [17,18].

Let f denote features, and $y \in \{1, \ldots, C\}$ to denote a class label, then the conditional probability of y given f is defined as as a soft-max over linear score functions:

$$p(y = c|f) = \frac{\exp(w_c^\top f)}{\sum_{c'=1}^{C} \exp(w_{c'}^\top f)}. \tag{4}$$

The likelihood is combined with a sparsity promoting Laplace prior over the parameters: $p(w) \propto \exp(-\lambda \|w\|_1)$, where $\|\cdot\|_1$ denotes the L_1 norm, and λ is set by cross-validation. Note that we can fix $w_1 = 0$ without loss of generality.

To learn the weight vectors we explore two strategies to use the set of noisy positive examples in the query set. The first is to simply treat all faces in the query set as positive examples ($y = 2$), and to use a random sample of faces from the database as negative examples ($y = 1$). The second strategy takes into account that each image in the query may contain at most one face of the queried person. To do this we learn the classifier iteratively, starting with all faces in query set as positive examples, and at each stage transferring the faces that are least likely to be the queried person from the positive to the negative set. At each iteration we transfer a fixed number of faces, which could involve several faces from a document as long as there remains at least one face from each document in the positive set. The last condition is necessary to avoid that a trivial classifier will be learned that classifies all faces as negative.

Once the classifier weights have been learned, we score the $(F + 1)$ assignments with the log-probability of the corresponding classifier responses, e.g. for $\gamma = 1$ the score would be $\ln p(y_1 = 2|f_1) + \sum_{i=2}^{F} \ln p(y_i = 1|f_i)$.

4 Using Query Expansion to Enhance Basic Filtering Methods

In this section we consider how query expansion can improve the results of the methods presented in the previous section. The original query is expanded as follows. First, we apply a named entity detector to the captions of the documents found using the initial text-based query, and count the number of occurrences of all names. The 15 most frequently detected names, we refer to these names as 'friends', will be used to query the database again. For each friend, we use only the documents in which the queried person does not appear in the caption, and if there are less than five such documents we discard the documents for the friend all together. This condition ensures a minimum amount of data from which we learn the model for each friend. We consider two ways to exploit the query expansion: (i) using the faces from the query expansion as a batch, and (ii)

using the described basic methods to process the faces returned on the query for each friend.

4.1 Query Expansion for Gaussian Mixture Filtering

The basic Gaussian mixture method uses only the faces in the query set. The first way to use the query expansion keeps the model as it is, but fits the 'background' Gaussian to the query expansion instead of the query set. In this manner the background Gaussian will be biased towards the friends of the queried person, and the foreground Gaussian is less likely to lock into one of the friends.

To use the query expansion in a more precise manner, we first apply the basic model to the query results of each friend. We then combine the foreground Gaussian of each friend with a generic background Gaussian in a mixture to form a more detailed query-specific background model. Thus, the difference with the basic model is that we set

$$p_{\mathrm{BG}}(f) = \frac{1}{N+1} \sum_{n=0}^{N} \mathcal{N}\left(f; \mu_n, \Sigma_n\right), \tag{5}$$

where $n = 0$ refers to the generic background model, and $n = 1, \ldots, N$ refer to the foreground models learned for friends 1 up to N. As before, we then fix p_{BG} and run the EM algorithm to find p_{FG} and the most likely assignment γ in each document.

4.2 Query Expansion for Linear Discriminant Filtering

The basic linear discriminant method uses a random sample from the database as negative examples to discriminate from the (noisy) positive examples in the query set. Our first way to use the query expansion is to replace this random sample with faces found when querying for friends. When there are not enough faces in the expansion (we require at least as many faces as the dimensionality to avoid trivial separation of the classes), we use additional randomly selected faces.

To use query expansion in a more precise manner we first use the basic linear discriminant method to learn a classifier to separate each friend $n = 1, \ldots, N$ from randomly selected faces, and similarly for the queried person. Let w_n denote the weight vectors learned for the friends, w_{FG} the weight vector learned for the queried person, and $w_{\mathrm{BG}} = 0$ the weight vector for the negative class. We then combine these weight vectors into a new multi-class classifier. Using this new multi-class classifier the probability that a face corresponds to the queried person is given by

$$p(y = \mathrm{FG}|f) = \frac{\exp(w_{\mathrm{FG}}^{\top} f)}{\exp(w_{\mathrm{FG}}^{\top} f) + \exp(w_{\mathrm{BG}}^{\top} f) + \sum_{n=1}^{N} \exp(w_n^{\top} f)}, \tag{6}$$

and similarly for the friends ($y \in \{1, \ldots, N\}$), and the background label ($y = \mathrm{BG}$). Note that in this manner the likelihood ratio between the background label and the queried person $p(y = \mathrm{BG}|f)/p(y = \mathrm{FG}|f)$ remains unchanged, and the same holds for the likelihood ratio of the background versus a friend. But, as desired, faces that are likely to be a friend get a lower probability to represent the queried person.

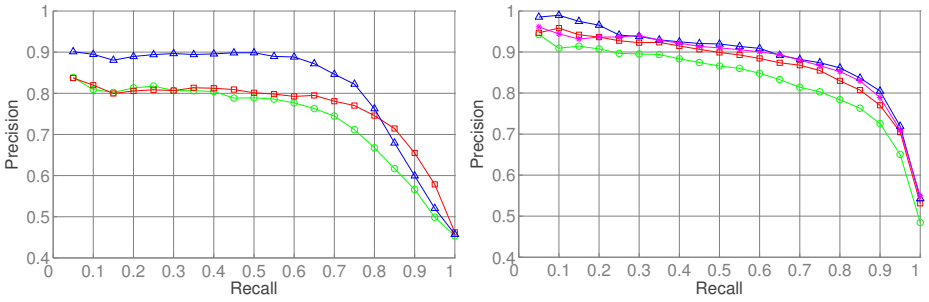


Fig. 3. Left panel: generative model using the query set to fit the background model (green, circles), the query expansion to fit the background (red, squares), and modeling each friend (blue, triangles). Right Panel: SMLR using the query set as positive examples (green circles), with iterative clean-up of the query set (red squares), using query expansion in the latter (blue triangles), and with the multi-class model (purple stars).

5 Experimental Results

We present results of two experiments. In the first experiment we evaluate performance of our different methods with and without query expansion using a database of captioned images downloaded from *Yahoo! News*. In the second experiment we consider whether the learned models are also useful to classify faces in the absence of captions.

5.1 Performance Evaluation of People Search based on Captions

Data Set and Pre-processing Pipeline. We evaluate the results of searches performed for the same 23 people also used in [12,13] over a collection of about 15.000 captioned news photographs downloaded from *Yahoo! News* in 2002–2003.[1] For each query we manually labeled all faces detected in the images returned on the text-based query.

On the image side, we used an off-the-shelf face detector [19], and facial feature detector [9]. The facial feature detector locates nine points on the face, and another four are determined from them. Before extracting features, each face image is filtered to suppress noise and to compensate for low-frequency lighting variations using a difference-of-Gaussian filter that has been reported to yield significantly better results for a collection of features [20]. For each of the 13 points on the face we calculate a 128 dimensional SIFT descriptor [21], yielding a $13 \times 128 = 1664$ feature vector for each face. On the caption side, we use also use an of-the-shelf named entity detector [22] to find the 'friends' for the query extension. The initial text-based query is solved by a simple string matching against all captions.

Experimental Setup. We test the two basic methods described in Section 3 in three settings: (i) not using query expansion, (ii) using the query expansion as a batch, and (iii)

[1] We would like to thank Tamara Berg for sharing the data set to us.

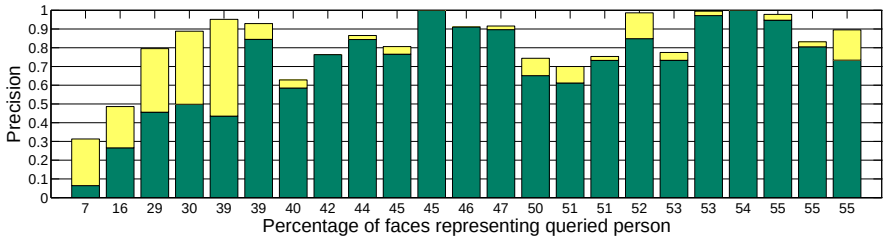


Fig. 4. Precision for the 23 queries obtained with the generative model, when using the query to fit the background model (green bars), and precision increase when using query expansion with a Gaussian for each friend (yellow bars). Queries are sorted by the percentage of faces in the query set that represent the queried person (see hor. axis).

using the query expansion while including processing of it using the basic method. We evaluate performance in terms of precision and recall measured over the faces detected in images returned on the text-based query. Precision-recall curves are averages over the curves obtained for individual queries: at a given level of recall we average over the 23 precisions.

Experimental Results. The precision-recall curves obtained using the generative model are presented in the left panel of Figure 3. When using a single Gaussian to model the background, fitting it to a random set of faces from the database or to the faces in the query expansion leads to very similar results. For clarity we omitted the result using random faces from the figure. An overview of precisions obtained for different people at 75% recall on average is is given in Figure 4. See the figure captions for more detail.

From the results we can draw the following conclusions. First, if we use a single Gaussian to model the background, then using the query set is a suboptimal choice: better performance is obtained when it is fitted to a random set of faces or to the query expansion. Second, the query expansion leads to substantial performance increases when we fit a Gaussian to each friend using the basic method: we observe three cases where precision is increased by more than 30%. In accordance with our goal, improvements are largest for people that have a low frequency among the faces in the query set.

The right panel of Figure 3 shows results obtained using the linear discriminant method, see the figure caption for details. Experimentally, we found values of $\lambda \in [5, 15]$ to give good results. This leads to sparse classifiers with on average 102 non-zero values in each weight vector (out of 1664) in the multi-class model learned on the expansions. The results lead to the following conclusions. First, iterative transfer of faces from the positive to the negative set leads to substantially better results. Second, query expansion improves results further by up to 4%. Third, the multi-class use of the query expansion does not lead to further improvements, in fact for many levels of recall it leads to a small drop in performance.

Fig. 5. Faces in the query set for Vladimir Putin (78% precision, 91% recall), assigned to the FG class (left), the BG class (middle), and classes associated with 'friends': Silvio Berlusconi, Hu Jintao, Sergei Ivanov, Jacques Chirac (from top to bottom, right).

In Figure 5 we show the result obtained using the multi-class model for a query for Vladimir Putin. For a collection of faces from the query set the break-down over different classes is shown. Note that not all expansion classes are pure, in the sense that only faces of one 'friend' are associated with them. However it is not critical for good retrieval performance how faces of other people are distributed over the BG class, and classes associated with other names: the only requirement is to make a good separation between the queried person and other people, not to recognize those related people.

We see that for both methods best results are obtained when using query expansion. We find the number of friends in the expansion not to have a significant impact on results once more than five are used.

In Figure 6 we show a summary of the best results obtained with the proposed methods, as well as the results obtained with the graph based methods of [12,13]. From the latter we consider the method using k-Nearest Neighbour graphs based on the number of matches between the same facial feature detections used by our generative and discriminative methods. We see that our discriminative method clearly outperforms the graph-based methods, and that our generative method with query expansion performs better than the graph-based methods up to about 85% recall. The results of the graph-

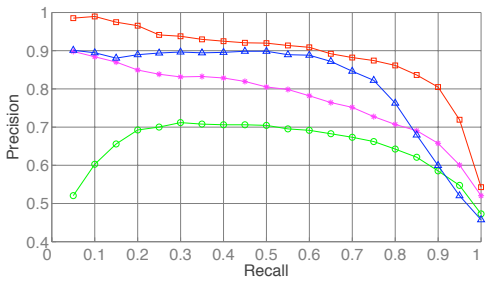


Fig. 6. An overview of the best results obtained with our discriminative model (red squares), our generative model (blue triangles), and the graph-based methods from Guillaumin et al. [13] (purple stars), and Ozkan and Duygulu [12] (green circles).

based methods are slightly different from those reported in [13] because before running these methods again, we removed several duplicate images from the data set and corrected some errors in the ground truth labeling.

Comparing the results obtained with the generative and discriminative method we note an interesting difference in the condition under which query expansion becomes most useful. For the former we need to model each friend in the expansion separately, and little is gained by fitting one simple Gaussian model to all the expansion data. For the latter the situation is reversed: using all faces in the query expansion as a bulk leads to a substantial gain, while modelling each friend separately does lead to further improvement. This difference can be understood as follows: a single Gaussian is unlikely to be a precise generative model as the expansion will contain faces of many different people, and it is unlikely that this will separate the queried person better than using a generic background model. Although the discriminative model is also limited, in this case we explicitly search for a linear separation of the query set and the expansion, rather than hope that the generative model will lead to good separation.

A second interesting observation is that without query expansion the discriminative model performs substantially better than the generative one (increases in the range $10-20\%$). However, when query expansion is used this difference becomes much smaller: less than 5% precision for recall in the range $25-75\%$. The generative model may therefore be preferred from a practical point of view when operating at these recall levels, as it is much faster to train using an efficient EM algorithm.

5.2 Classifying Faces without Accompanying Captions

Once a model has been learned from a database of captioned news images, we can use the model also to find the person in other images even if they do not have a caption. To test this we have learned a linear discriminant model for ten people from the *Yahoo! News* database, and then apply the model to images of the *Labeled Faces in the Wild* database [23]. This database contains about 13.233 images of people cropped to head-and-shoulders, the faces are labeled with names, but captions are not available. The

Fig. 7. For each query the first ten images returned from *Labeled Faces in the Wild*.

database is also gathered from material of Berg et.al. [14], but includes a ground-truth labeling of all faces, and allows us to test our models on faces that were not in the result of a text-based query. To allow unambiguous use of the labels, we selected the 11.948 images for which our face detector found exactly one face; in 289 images no face was detected, and in 996 more than one face was detected.

We use trained discriminative models to classify all faces, and then calculated precision at three levels of recall, averaged over the ten persons. As before, the precisions of the multi-class and two-class models were found to be very similar: here differences were smaller than 1%. The precision among the first 10 faces was 99% (1 error), degrading to 92% at 25 faces, and to 67% at 100% recall. These precisions are comparable to those measured on the query set from the *Yahoo! News* database. This is encourag-

ing because in the current setting many more negative images are present in the test set. Figure 7 shows the first ten images for these ten queries.

6 Conclusions

We have shown how query expansion leads to improved results when searching for people in captioned news images. Although queries for which text-based search in the caption leads to a low fraction of relevant faces remain difficult, we have made significant progress in these cases, boosting precision by 20% up to 50% for the generative model in the five most difficult cases.

Query expansion is particularly useful for the Gaussian mixture approach to filter the initial query result. Without query expansion the discriminative method clearly outperformed the generative one ($\sim$ 10% difference in precision), but these differences become smaller using query expansion ($\sim$ 5% differences in precision).

We achieve performance levels that are significantly higher than those obtained in [12,13] using similarity-based methods. We obtain our results with methods that do not require calculation of pairwise similarities and are therefore much faster when many faces are processed. Our best method (SMLR + expansion) obtains a precision of 87.4% (83.6%) for a recall of 75% (85%), while the best previously reported result on these queries only reaches 77.6% (73.0%) for the same recall values. It is not obvious how query expansions can be used in a similarity-based approach, as the graph-density score takes into account only the similarities between faces selected for the queried person, and not those between selected and non-selected faces.

In our final experiment we have shown that the learned models also perform well when classifying faces without captions. This suggests that once a model has been learned from caption-based supervision, we can also identify these people in images where they are not named in the caption.

A demo of the generative model with query expansion on the *Yahoo! News* data set is available at `http://lear.inrialpes.fr/~verbeek/facefinder`.

References

1. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12) (2000) 1349–1380
2. Ponce, J., Berg, T., Everingham, M., Forsyth, D., Hebert, M., Lazebnik, S., Marszalek, M., Schmid, C., Russell, B., Torralba, A., Williams, C., Zhang, J., Zisserman, A.: Selected Proceedings of the first PASCAL Challenges Workshop. Lecture Notes In Artificial Intelligence. In: The 2005 PASCAL Visual Object Classes Challenge. Springer (2006)
3. Lazebnik, S., Schmid, C., Ponce, J.: Affine-invariant local descriptors and neighborhood statistics for texture recognition. In: Proceedings of the IEEE International Conference on Computer Vision. (2003) 649–655
4. Verbeek, J., Triggs, B.: Region classification with Markov field aspect models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2007)
5. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., Jordan, M.: Matching words and pictures. Journal of Machine Learning Research **3** (2003) 1107–1135

6. Grangier, D., Monay, F., Bengio, S.: A discriminative approach for the retrieval of images from text queries. In: Proceedings of the European Conference on Machine Learning. (2006) 162–173
7. Bressan, M., Csurka, G., Hoppenot, Y., Renders, J.M.: Travel blog assistant system. In: Proceedings of the International Conference on Computer Vision Theory and Applications. (2008)
8. Jain, V., Learned-Miller, E., McCallum, A.: People-LDA: Anchoring topics to people using face recognition. In: Proceedings of the IEEE International Conference on Computer Vision. (2007)
9. Everingham, M., Sivic, J., Zisserman, A.: 'Hello! My name is... Buffy' - automatic naming of characters in TV video. In: Proceedings of the British Machine Vision Conference. (2006) 889–908
10. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2008)
11. Marcel, S., Abbet, P., Guillemot, M.: Google portrait. Technical Report IDIAP-COM-07-07, IDIAP (2007)
12. Ozkan, D., Duygulu, P.: A graph based approach for naming faces in news photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2006) 1477–1482
13. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Automatic face naming with caption-based supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2008)
14. Berg, T., Berg, A., Edwards, J., Maire, M., White, R., Teh, Y., Learned-Miller, E., Forsyth, D.: Names and faces in the news. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2004) 848–854
15. Buckley, C., Salton, G., Allan, J., Singhal, A.: Automatic query expansion using SMART: TREC 3. In: Proceedings of the Text Retrieval Conference. (1995) 69–80
16. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: Proceedings of the IEEE International Conference on Computer Vision. (2007)
17. Krishnapuram, B., Carin, L., Figueiredo, M., Hartemink, A.: Sparse multinomial logistic regression: fast algorithms and generalization bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(6) (2005) 957–968
18. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2007)
19. Mikolajczyk, K., Schmid, C., Zisserman, A.: Human detection based on a probabilistic assembly of robust part detectors. In: Proceedings of the European Conference on Computer Vision. (2004) 69–81
20. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. In: Analysis and Modelling of Faces and Gestures. Volume 4778 of LNCS., Springer (oct 2007) 168–182
21. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
22. Deschacht, K., Moens, M.: Efficient hierarchical entity classification using conditional random fields. In: Proceedings of Workshop on Ontology Learning and Population. (2006)
23. Huang, G., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (2007)