2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology

Automatic Face Annotation in News Images by Mining the Web

Eric Medvet, Alberto Bartoli, Giorgio Davanzo, Andrea De Lorenzo
DIII, University of Trieste
Via Valerio, Trieste, Italy
emedvet@units.it

***Abstract*—We consider the automatic annotation of faces of people mentioned in news. News stories provide a constant flow of potentially useful image indexing information, due to their huge diffusion on the web and to the involvement of human operators in selecting relevant images for the stories. In this work we investigate the possibility of actually exploiting this wealth of information.**

We propose and evaluate a system for automatic face annotation of image news that is fully unsupervised and does not require any prior knowledge about topic or people involved. Key feature of our proposal is that it attempts to identify the essential piece of information—how a person with a given name looks like—by querying popular image search engines. Mining the web allows overcoming intrinsic limitations of approaches built above a predefined collection of stories: our system can potentially annotate people never handled before since its knowledge base is constantly expanded, as long as search engines keep on indexing the web. On the other hand, leveraging on image search engines forces to cope with the substantial amount of noise in search engine results. Our contribution shows experimentally that automatic face annotation may indeed be achieved based entirely on knowledge that lives in the web.

***Keywords*-face recognition, image annotation, web mining, SURF**

I. Introduction

Automatic image annotation is an important and challenging task in many practical applications. Although solving this problem in its more general form requires sophisticated image understanding capabilities, significant results may usually be obtained by simple image-based analysis coupled with proper exploitation of contextual information, e.g., the textual content of a web page in which the image is found. From this point of view, *news stories* constitute an increasingly important source of potentially useful information. On the one hand, the web is an extremely prolific producer of news stories composed of textual content coupled with one or more images. On the other hand, the content of those images has been deemed appropriate to the text by a human operator. In this work we examine the possibility of exploiting this wealth of information. We propose and evaluate an unsupervised method for automatic annotation of images in terms of the *names* of the people involved in the story.

Key feature of our approach is that it does not require any previous knowledge about text or people involved in the story; nor does it require any prior examples of sample news stories. Instead, our approach exploits the knowledge which lives in the web by querying image search engines—Google Images, Bing Images and Yahoo! Image Search. Our system takes a story, i.e., a pair text-images, and works as follows: (i) extracts all the person names from the story text; (ii) extracts all the faces from the images and, (iii) for each person name, associates the name with the correct face. The main contribution of our proposal is in step iii, which is performed by querying popular image search engines for the given person name and figuring out which faces, if any, are similar to the results. In other words, we attempt to identify the information of interest—the aspect of a person whose name is known—by mining the huge yet noisy knowledge embedded in these engines.

Mining image search engines has several major advantages over usage of a predefined collection of stories as a training set, even if training images have not been annotated. First, the system need not be prepared to cope with people never handled before. If the system analyzes a story where a person P is found that was never encountered by the system before (and hence his face is completely unknown to the system), it may nevertheless be the case that images of P exist somewhere in the web and have been indexed by image search engines. Second, the system is constantly updated as long as image search engines keep on indexing the web. The system may thus annotate faces of people which become suddenly famous—players of the recent Fifa World Cup, for example. Third, image search engines are robust to imprecise queries. This property is essential for handling references to people in textual news effectively—for example, news may mention either the name “Barack Obama” or the epithet “President Obama” which is not the actual name of any person.

Relying on image search engines is not necessarily feasible, though. Leveraging their results for coupling a person name with a face automatically could be possible to a limited extent or not at all. It is often the case that a single

978-0-7695-4513-4/11 $26.00 © 2011 IEEE
DOI 10.1109/WI-IAT.2011.101

image result represents two or more people, without any hint about which one corresponds to the queried person. Besides, there is a large variation of pose, illumination conditions, occlusion and facial expression in the image results, even when they all contain only the queried person. Last but not least, images that are not relevant or that do not correspond to the queried person often appear in the results. Our contribution includes the experimental assessment that web mining is indeed suitable to the specific problem.

II. Related work

The application of face recognition techniques to the problem of automatic annotation of news—or of captioned images—has gained large interest in recent years [1], [2], [3], [4], [5], [6]. This problem is particularly challenging because it involves faces captured in broadly varying configurations with respect to viewpoint, expression, illumination and so on. All approaches consider a database of stories and then apply some sort of clustering in order to annotate all the faces that appear in at least a given number of story images within the database. In other words, person names found in the text—or caption—accompanying an image are considered as weak labels which have to be refined later in order to become annotations. Such approaches are fully unsupervised but they can handle only people indeed present in the database. Moreover, they can hardly annotate people who rarely appear in the database. In our approach we investigate the possibility of using the Web as source of knowledge for recognizing people, thereby leveraging the the coverage of image search engines while remaining fully unsupervised.

A method for clustering and annotating face images in captioned news images is proposed in [1]. The method builds clusters of faces of the same person based on a modified version of the Eigenfaces [7] approach, i.e., kernel Principal Component Analysis (kPCA) and Linear Discriminant Analysis. Clusters are then annotated with the corresponding person name based on the captions. Similarly to us, extraction of names from captions consists of an heuristics method which considers two or more capitalized words followed by a present tense verb. The authors of this study further extended their work by adding a context-based language model which allow them to improve clustering accuracy [8].

A similar approach is proposed by [3]. This work focuses on Japanese news, hence a Japanese morphological analyzer is used in order to extract person names from text. Clustering of faces of the same person is based on the Eigenfaces method, which is applied according to an iterative k-means clustering procedure. Extraction of faces from the news image relies on an AdaBoost [9] face detector, with a bag-of-keypoints enhancement which is proposed in order to improve precision. In our experiments we found that the precision of the AdaBoost face detector was sufficiently high to our purposes and hence we used it without any enhancement.

The problem of annotating all faces in a database of news, as well as finding all faces in the database of a named person, is considered in [2]. A single method for both flavors of the problem is proposed, which bases on a graph connecting faces: the higher the similarity between faces, the stronger the connection; clusters are then built taking into account densest components of the graph. Similarly to our work, they propose a metric for face similarity which considers matching interest points. Their metric is based on SIFT descriptors obtained at 13 facial features [10] (an improvement of the use of SIFT descriptors has been proposed later in [11].), whereas we use SURF interest points and descriptors [12]. Speeded-Up Robust Features (SURF) have been recently experimented as an effective approach in face recognition tasks [13].

Other graph based approaches for annotating faces with names in news have been presented in [5] and [6]. The authors of the former use SIFT features to construct a similarity graph of faces in a small training set; they then annotate unknown faces with the nearest known face node or with the closest category, being a category a set of faces of the same person. In the latter, a graph is built, basing on image search engine results, for each person name in the news; then the unknown faces are associated with the nearest dense graph.

A probabilistic approach is followed in [4]. The authors propose a model which describes how names in the text generate faces in the image and another model which describes how faces generate names; then, they propose a third model which computes the linking between names and faces by considering their joint probability. The Expectation Maximization algorithm is used to tune the parameters of the models. Face extraction and clustering is again based on AdaBoost face detector, followed by the computation of a face descriptor based on texture and shape of a number of facial features extracted according to [10]. Extraction of person names from captions is performed with a named entity recognizer based on maximum entropy. Disambiguation of variants of names that refer to the same person uses a dictionary. We do not need to perform disambiguation of person names because we exploit the robustness of search engines in this sense.

A slightly different but related problem is considered in [14], which proposes a method for associating topics with people based on face recognition. Since this approach is focused on topics rather than on person names, captions are not analyzed searching for candidate person names. Indeed the presence of a person name in a caption is not even a constraint. Building of topics is based on a hyper-feature based face identifier: hyper-features are basic features (positions, intensity values and edge energies in different directions) which are used to decide which image patches are most

relevant for identifying a given object class. Two interesting applications concerning the name-face annotation problem are given by [15] and [16], which consider faces appearing in video sequences provided with weak labels.

A method for retrieving relevant faces of one person out of search engines results is proposed in [17]. Although this work is not concerned with automatic annotation of captioned images, some of the crucial issues in terms of noise accompanying images are very similar. First, the presence of irrelevant faces in the image results returned by the engine. Second, the wide variability in image face conditions—pose, illuminations, expression and so on. An iterative procedure is proposed to automatically classify search engine results as relevant or irrelevant. The procedure consists of two steps: in the first step ranking is computed using the Local Density Score—as we do—which is then used in the second step to train a set of weak classifiers whose output is combined in a bagging-based framework. The initial ranking is computed using the Eigenfaces approach on the set of results returned by the queried search engine. The concept of re-ranking image search engine results is explored also by [18], where the authors use the text of the original page from which the image was obtained in order to assess its relevance.

III. The proposed system

Our system processes a story, that is a pair composed of a text t and a set of associated images $i_1, i_2, \ldots$. The result of the processing consists in the annotation of faces detected within $i_1, i_2, \ldots$ based on the person names found in t.

The system works as follows. First, the text t is analyzed and a set of possible person names $n_1, n_2, \ldots$ is extracted. If no name can be extracted, the processing ends. The details about this step are described in Section III-A.

Second, each image i of the story is analyzed and a set of faces is extracted from the image. The set $f_1, f_2, \ldots$ denotes the faces extracted from all the story images. The details about this step are described in Section III-B.

Third, for each name n and each face f a procedure is executed in order to compute a *dis-coupling score* for the pair n, f. If the dis-coupling score is lower than a fixed threshold, then the face f is annotated with the name n. This *name-face coupling* procedure is our main contribute and actually involves mining the image search engines in order to figure out relevant annotations for the face. The details about this step are described in Section III-C.

Figure 1 summarizes graphically the processing flow of our system.

A. Names extraction

We use a composite approach for extracting person names from a text t, a name being composed of one or more words. First, we use a named entity recognizer which is based on a maximum entropy classifier from the OpenNLP[1] package:

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

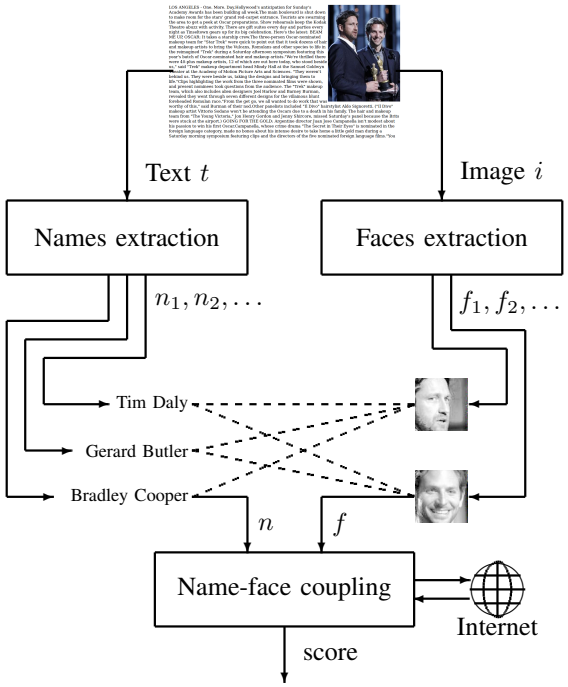


Figure 1. The processing flow of the proposed system.

in this way, we obtain a set of candidate person names $N' = \{n'_1, n'_2 \ldots\}$.

Second, we build another set of candidate person names $N'' = \{n''_1, n''_2, \ldots\}$ by identifying all sequences of at least 2 and at most 5 capitalized words within t: we retain all and only the word sequences in which at least one word is contained in a set of possible person first names. We compiled such set a priori by merging two lists of Italian and English names available on the web.

Finally, we construct a single set of candidate names N by merging N' and N'', i.e., $N = N' \cup N''$.

We also explored the possibility of identifying names within the story text by means of keywords extraction. We thought that looking for keywords could provide more precise results, which should also be more similar to textual hints that search engine actually use while indexing images. We performed several experiments with the Yahoo[2] and Zemanta[3] term extraction APIs, but we obtained a far lower recall and a comparable precision with respect to the approach described above.

B. Faces extraction

We describe here how we extract a set of faces from a single image. We repeat the procedure for each image of the story. The face extraction procedure consists of three steps: detection, normalization and projection. The

[2]http://developer.yahoo.com
[3]http://developer.yahoo.com

procedure output is a set of numeric vectors $f_1, f_2, \dots$ which describes the faces: we call f_i a *face vector*.

First, we convert the image to gray-scale and use the AdaBoost face detector module [9] included in the Open Computer Vision Library (OpenCV) in order to extract the rectangular portions of the image containing human faces.

Then, we use the eye detector module included in OpenCV in order to extract the positions of subject eyes within the face image. If this module finds exactly two eyes with sufficient confidence, we use their positions to apply a face image rotation in order to obtain a canonical pose for the subject; otherwise, we leave the face image as found by the face detector. We crop the face image to the size of 100×100 pixel and perform an histogram equalization.

The final step is projection: the goal of this step is to obtain a numeric representation f of the face image. To this end, we performed two alternative kinds of elaborations. One consists on a space dimension reduction based on the Principal Component Analysis and is described in Section III-B1. The other is based on the extraction of a set of interest points on the image and on the numerical representations of these points: this technique is described in Section III-B2. Section IV provides insights about the effectiveness of the two methods.

1) PCA projection: This projection method bases on the Eigenface representation [7], which describes a face as a gray-scale vector compressed by means of Principal Component Analysis (PCA).

We consider a point $f' \in [0,1]^{100\times100}$ which corresponds to the gray-scale representation of the face image. Then we project f' onto the eigen-subspace we obtained a priori by performing PCA on a set of 500 face images. We obtained such face images by randomly extracting a number of images from news web sites; we performed on them the detection and normalization steps described above. Then, we chose the 128 principal components corresponding to the largest eigenvalues, resulting in a cumulative percentage of the sum of eigenvalues—i.e., the portion of variance—of 90%.

We performed a set of preliminary experiments and found that the number of faces chosen to perform PCA as well as the portion of variance taken into account are not sensitive parameters.

The numeric representation f of the given face image is hence a vector of size 128.

2) SURF projection: The rationale of this projection method is to extract a set of *interest points* on the face image, each one being located on the image in a point interesting from the visual point of view. Then, the numerical *descriptions* of the interest points are used together as a description for the whole image.

We use a technique called SURF (Speeded-Up Robust Features) [12] for both the interest point detection and description steps. In the detection step, a set of blob-like structures is chosen on the image as interest points at locations where the determinant of the Hessian matrix is maximum. Then, a 64-dimensional descriptor is computed for each interest point: the descriptor focuses on the spatial distribution of gradient information within the interest point neighborhood.

In detail, given the input face image, we detect the interest points: the number of interest points actually detected in a given face image is variable and depends on the image itself. We found that about 130–170 interest points are extracted from each face image. For each interest point we obtain a 64-dimensional U-SURF descriptor s and the coordinates p of the interest point in the image. The numeric representation of a face image is hence given by $f = \langle \{s_1, s_2, \dots\}, \{p_1, p_2, \dots\} \rangle$.

SURF is robust to perturbation on image scale, rotation, brightness and contrast and to some degree of noise. SURF is not specifically tailored to face images but has been recently used in the scenario of face recognition [13] and proved to be faster than the widely used SIFT approach [19] while achieving not worse error rates. As done in [13], we actually used the upright variant of SURF (i.e., U-SURF) that is not fully invariant with respect to the image rotation, but still maintains a robustness to rotations of $\pm 15°$, which is sufficient in the context of face recognition. U-SURF is also faster to compute and can increases distinctivity with respect to SURF [20].

C. Name-face coupling

The goal of the name-face coupling procedure is to indicate whether a given name n is a suitable annotation for the face vector f, in other words, if n is a name for the person represented by f.

We proceed as follows. We submit a search query composed of the text n (lowercased) to three different image search engines: Google Images, Bing Images and Yahoo! Image Search. For the two former search engines, we enable the "face" option which makes the search engine prefer results that include a human face within the image. We hence obtain three sets of images that we merge forming a single set I.

Then, we apply the face extraction procedure described in Section III-B to each image of I and obtain a list of face vectors F. Note that a single image of I could correspond to zero, one or more face vectors of F; we discard exact duplicate face vectors while building F. We sort F accordingly to the ranking of the result from which each F element came from: e.g., f_1 is the face extracted from the first result from Google Images, f_2 is the face extracted from the first result from Bing Images, and so on. Finally, we truncate the list F to its l first elements.

At the end we compute the dis-coupling score for the pair f, n, for quantifying the relevance of the face described by f with respect to the query results $F = \{f_1, f_2, \dots\}$: the lower the score, the greater the relevance of f with

respect to the query results and, hence, with the name n. For the dis-coupling score, we use the idea of density-based clustering [21], [22], borrowing from [17]: the Local Density Score (LDS) of a face vector f is defined as the average distance to its k-nearest neighbors:

$$\mathrm{LDS}(f) = \frac{\sum_{f_j \in R_k(f)} d(f, f_j)}{k}$$

where $d(f, f_j)$ is the distance between f and f_j and $R_k(f)$ is the set of the k-nearest neighbors from f, according to the given distance. We compare $\mathrm{LDS}(f)$ with a fixed threshold τ: if and only if the score is lower than τ, we annotate the face given by f with the name n.

Concerning the distance $d(f, f_j)$ we used two methods, depending on the projection method used in the face extraction step.

For the PCA projection method, we simply compute the Euclidean distance between the two vectors. We also explored another distance based on the number of shared neighbors between two elements (which itself bases on the Euclidean distance), as done in [17], but we found by preliminary experimentation that it was far less effective.

For the SURF projection method, the distance between two projected faces f and f' is defined as $d(f, f') = \frac{1}{m+1}$, where m is the number of *matching interest points* between f and f'. We consider that two interest points match if their visual appearance is sufficiently similar, i.e., if the distance between the corresponding descriptors is low. To this end, we compare each descriptor s_i of f to the descriptors of f': if the Euclidean distance of s_i from the nearest descriptor s'_j is less than α times the Euclidean distance from the second nearest descriptor of f', then we say that s_i and s'_j match. We set $\alpha = 0.5$ accordingly to [13], which also showed that it is not a sensitive parameter.

As an optimization, we actually compare each descriptor of f with only a subset of the descriptors of f'. Since the two face images are obtained in the same way—using the AdaBoost face detector and then performing a normalization step—they likely have a similar viewpoint. Hence, a convenient choice is to compare each interest point of one face only with the interest points of the other face which lie in the same face zone: e.g., an interest point corresponding to a corner of the former face right eye will be compared only with interest points which lie in the second face portion where the right eye likely lies too. Accordingly with this choice, we compare s_i of f only to the descriptors s'_j of f' for which $(x_i - x'_j)^2 + (y_i - y'_j)^2 \leq r$ where r is a given distance threshold between the two interest points. A viewpoint consistency constraint is hence imposed. The lower the parameter r, the tighter the constraint; the greater the parameter r, the larger deformation of the second face image is allowed, while searching for matching interest points.

In summary, the name-face coupling procedure uses the following 4 parameters:

- l the size of the truncated list F of faces obtained from search engines;
- k the number of neighbors in the LDS computation;
- τ the threshold for the LDS score;
- $d(\cdot, \cdot)$ the distance used for the LDS computation: in particular we choose one among Euclidean distance between PCA projections and the SURF matching distance, which bases on the parameter r.

In Section IV-A we present an experimental evaluation of the name-face coupling procedure which also assesses how these parameters affect the procedure effectiveness.

IV. Experimental evaluation

We present here the results of a set of experiments we performed in order to assess the effectiveness of our system. We first show the experimental evaluation results of the three steps—names extraction, faces extraction, name-face coupling—each taken alone; then we show the result of the full system evaluation.

We used two different dataset for our experiments. For the name-face coupling step, which is the most important component of the system, we build a dataset of 30 pairs of person names and corresponding face images: this dataset is further described in Section IV-A.

For the evaluation of names and faces extraction steps, as well for the full system evaluation, we compiled a dataset of 21 stories by randomly selecting recent news items from the US version of Yahoo news web site. We considered the following categories: U.S. National, Politics, Sports and Entertainment. Each story is composed by a piece of text and a single image.

We processed each story as described in the previous section; then, we manually checked the results. We extracted 174 names and 30 faces from the stories and we extracted about 87000 face images from image search engines results.

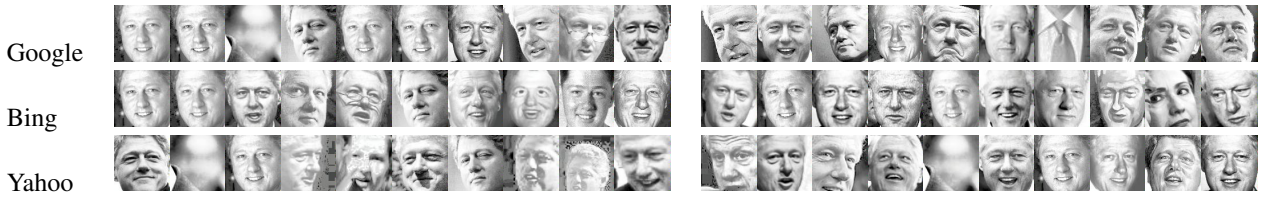
Concerning the names extraction step, we found a recall of 97% and a precision of 25%. As a comparison, on the same dataset we obtained a recall of 25% and a precision of 50% when extracting names by means of keywords extraction (see Section III-A).

Concerning the faces extraction step, we found a precision of 92% and a recall of 89%; most of the missed faces were indeed not in a frontal pose and hence difficult to detect with the AdaBoost face detector.

Experimental results for name-face coupling step are given in the next section and for the full system in Section IV-B.

A. Name-face coupling

In order to evaluate the name-face coupling procedure, we compiled a list of 30 person names. The list includes quite famous politicians (e.g., Eric Massa, Erik Paulsen),



(a) Search engines results for the query "Bill Clinton"

Figure 2. A subset of the face images extracted with our procedure (see Section III-B) from search engine results for the person name "Bill Clinton". Each row of images correspond to a search engine. The 10 face images on the left are extracted from the first 10 results; the 10 face images on the right are extracted from the 50th and subsequent results.

actors (e.g., Alexander Skarsgård, Christopher Walken) and sport players (e.g., Dino Zoff, David Villa). Figure 2 shows a subset of the face images extracted from search engine results for the person name "Bill Clinton". It can be seen that illumination and pose conditions vary over obtained face images.

For each name of the list, we collected one image representing the corresponding person alone and we obtained the corresponding face image by applying the face extraction procedure described in Section III-B. Then, again for each name of the list, we applied the name-face coupling procedure (a) to the name and the corresponding face image and (b) to the name and 5 other face images of different people of the list (chosen in a repeatable way).

We repeated the experiment with different values for the 3 parameters l, k and $d(\cdot,\cdot)$. For each combination, we measured the false positives rate (FPR) and false negatives rate (FNR) with different values for the threshold parameter τ, which enabled us to obtain the receiver operating characteristic (ROC curve). We experimented with 4 distances: the Euclidean distance applied on PCA projection (PCA, in brief) and three versions of the SURF matching distance with different values for r (i.e., the distance threshold for searching matching interest points). We denote the resulting versions as SURF-2, SURF-5 and SURF-10. We varied $l \in \{30, 50, 100, 200, 500\}$ and $k \in \{2, 3, 5, 10, 20\}$: these settings corresponded to about 15000 face images and require the computation of about 90000 distances between face images, once for each of the four considered distances.

Figure 3 shows the total error rate (i.e., FPR + FNR) vs. l, one curve for each distance, with a fixed value for $k = 5$. Each point represents the best point of the ROC which has been obtained with the given distance and l value.

First and foremost, this result suggests that our approach is indeed effective: with $l = 500$ and SURF-5, we obtain a total error rate lower than 0.04. It also shows that the name-face coupling procedure may indeed be practical, as the total error rate is moderate even with a small value for l (FPR + FNR ≈ 0.15 with $l = 50$). This is an interesting result because, depending on the actual deployment of the system, it could be convenient to elaborate only few search engine results.

Figure 3 shows that the SURF distance performs in general better than the PCA distance: in particular, for smaller r, SURF distance exhibits a total error rate which is about 0.2 lower than the one obtained with PCA. Another interesting finding concerns the impact of l: it can be seen that the larger the number of images obtained from search engines, the lower the error rate. This result is not surprising, because with more results it is likely to find one or more images of the queried person in which face condition are more similar to the face image under investigation.

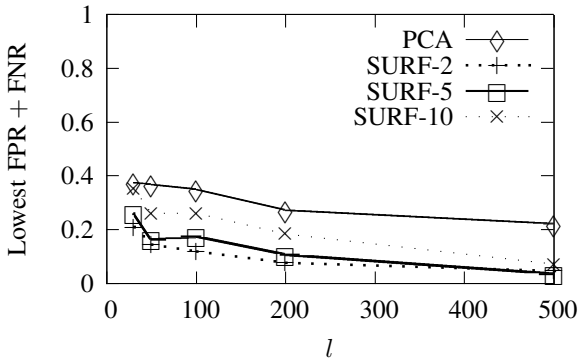


Figure 3. Lowest FPR + FNR at fixed $k = 5$ with different values for l. Each point represents the best point of the ROC in the given condition.

Figure 4 show the total error rate (i.e., FPR + FNR) vs. k, one curve for each distance, with fixed values for $l = 50$. From the experimental result, the impact of k on the name-face coupling procedure is less sharp than the one of l. In general, it can be seen that higher values for k correspond to greater error rate: yet, SURF distance appear to be quite robust to k, in particular for $l = 500$. It worths to note that values too low for k could lead to an exceeding high sensibility to noise within search engines results: e.g., if we couple the name "Bill Clinton" with a face image of George W. Bush with $k = 1$, a single result of the latter will lead to a false positive. In this sense, the robustness with respect to k is a plus for the SURF-2 and SURF-5 distances.

Basing on the above analysis concerning the impact of l and k, we chose the parameters value of $l = 500$ and $k = 5$. Figure 5 shows the ROC curves for this working point: it can be seen that SURF distance outperforms PCA distance, regardless of the value of the parameter r. SURF-5 distance obtains a FPR lower than 0.03 with FNR lower than 0.01,

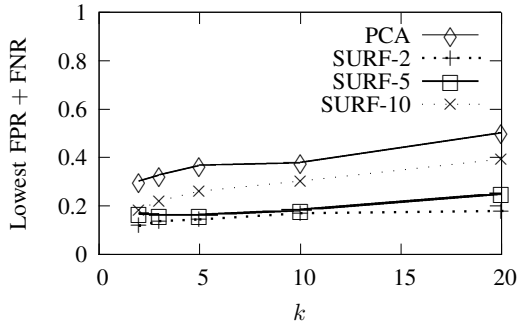


Figure 4. Lowest FPR + FNR at fixed l. Each point represents the best point of the ROC in the given condition.

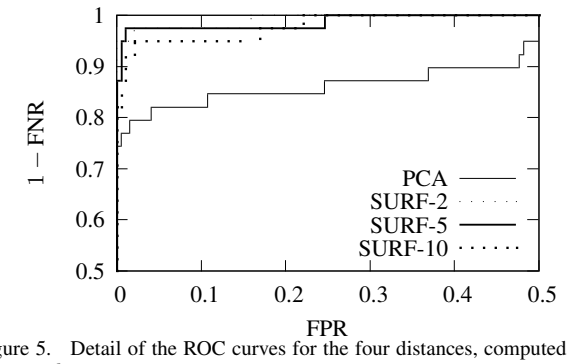


Figure 5. Detail of the ROC curves for the four distances, computed with $l = 500, k = 5$.

whereas in the best case PCA obtains a FPR + FNR greater than 0.2. These figures show that our name-face coupling procedure is effective in isolation.

Table I shows the total error rate for the best point in the ROC obtained with $l = 500$ and $k = 5$. Along with the numerical data concerning curves of Figure 5, which is shown in the first column, the table shows also the lowest FPR+FNR obtained while using only results of single search engines. These results confirm the intuition that combining the results from the three engines is indeed highly beneficial, irrespective of the similarity distance actually used. It can also be seen that there are not significant differences among the three search engines. It is interesting to observe that our name-face coupling procedure seems to work slightly better with Yahoo! Image Search results, despite the fact that this is the only image search engine which does not provide an option for filtering results basing on the presence of faces.

B. Full system

We fed our system with the stories composing our dataset (see Section IV). We used the optimal parameters configuration that we found in the experimental evaluation of the name-face coupling step: SURF-5 distance, $l = 500$ and $k = 5$.

Distance	All	Bing	Yahoo	Google
PCA	0.22	0.37	0.46	0.32
SURF-2	0.05	0.11	0.13	0.15
SURF-5	0.04	0.14	0.09	0.14
SURF-10	0.07	0.22	0.18	0.35

Table I
LOWEST FPR + FNR COMPUTED WITH $l = 500, k = 5$, FOR EACH OF THE FOUR DISTANCES. EACH COLUMN CORRESPONDS TO A DIFFERENT SEARCH ENGINE SETTING: ALL ENGINES OR EACH SINGLE ENGINE.

We manually checked the names the system applied to each face and found that our system obtains an accuracy of 71.2% and a recall of 70.4%. As a comparison, we also measured the effectiveness of our system using the PCA distance (with a suitable value for τ) and we found an accuracy of 61.9% and a recall of 47.6%.

Concerning the recall, which we assessed at 70.4%, we think that the result is promising and suggests that our system may be an effective aid in information retrieval tasks.

The accuracy figures are consistent with the results obtained by other approaches: 72% in [1], 71% in [4] and 63% in [2]. We remark that the cited works are based on a database of news collected in advance. It follows that they perform poorly on people who appear only rarely within the pre-built database and, obviously, cannot annotate people who never appear in the database. On the contrary, our approach can potentially annotate each person for which image search engines return a reasonably precise set of results. In other words, our approach inherits the coverage of image search engines and, as such, it may potentially start annotating people whose popularity suddenly arises, provided that search engines cover the corresponding event. This result, therefore, indicates that we can retain the accuracy of earlier approaches without being constrained by the closed world of a predefined and statically built news database.

V. CONCLUSIONS

We proposed and assessed experimentally a method for automatically annotating person faces which appear in news images. We extract the person names from the news text and the face images from the news image. Then, we attempt to associate each name with each face. Differently from other previously proposed approaches which rely on large databases of news, we do not require any previous knowledge about people. We can hence annotate also people never handled before or people whose popularity suddenly arises—e.g., medalists at the recent Winter Olympic Games.

We use the Internet as a live source of knowledge in order to decide which names, if any, should be associated with each face: we query three different image search engines—Google Images, Bing Images and Yahoo! Image Search—for a given name and then we obtain a similarity measure between the face under investigation and the faces extracted from query results. The face annotation task is made difficult by the large amount of noise, of various forms, widely present in news text, news images and results returned by image search engines.

Our experimental evaluation shows an accuracy in the order of 70%, a value that appears to be sufficiently high to be practically useful for the specific problem addressed.

Moreover, and most importantly, this figure is in line with similar experiments obtained on a predefined news database. It follows that the shifting from a closed world built around a predefined set of people, to an open and continuously expanding world of relevant people is indeed possible and may be obtained without necessarily sacrificing accuracy.

Our results suggest a promising approach for large-scale face annotation, which is entirely based on knowledge that lives in the web. On the one hand, it leverages the power, sophistication and coverage of modern image search engines. On the other hand, it exploits the wealth of image indexing information embedded in the constant flow of news stories produced every day.

REFERENCES

[1] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller, and D. A. Forsyth, "Names and faces in the news," in *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 2. Los Alamitos, CA, USA: IEEE Computer Society, 2004, pp. 848–854.

[2] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Automatic face naming with caption-based supervision," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1–8.

[3] A. Kitahara, T. Joutou, and K. Yanai, "Associating faces and names in japanese photo news articles on the web," in *Proceedings of the 22nd International Conference on Advanced Information Networking and Applications - Workshops*. IEEE Computer Society, 2008, pp. 1156–1161. [Online]. Available: http://portal.acm.org/citation.cfm?id=1395080.1395454

[4] M. F. Moens and T. Tuytelaars, "Linking names and faces: Seeing the problem in different ways," 2008.

[5] H. Zitouni, M. Bulut, and P. Duygulu, "Recognizing faces in news photographs on the web," in *Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on*, 2009, pp. 50–55.

[6] D. Ozkan and P. Duygulu, "Interesting faces: A graph-based approach for finding people in news," *Pattern Recognition*, vol. 43, no. 5, pp. 1717–1735, May 2010. [Online]. Available: http://www.sciencedirect.com/science/article/B6V14-4XSJVJ5-1/2/6595aaec3676a87d33f0b4d501191a8f

[7] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 591, 1991.

[8] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth, "Who's in the picture?" in *Advances in neural information processing systems 17: proceedings of the 2004 conference*. The MIT Press, 2005, p. 137.

[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 1. Los Alamitos, CA, USA: IEEE Computer Society, 2001, p. 511.

[10] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is... buffy–automatic naming of characters in TV video," *BMVC*, pp. 889–908, 2006.

[11] J. Sivic, M. Everingham, and A. Zisserman, ""Who are you?" - learning person specific classifiers from video," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 1145–1152.

[12] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-Up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, Jun. 2008. [Online]. Available: http://www.sciencedirect.com/science/article/B6WCX-4RC2S4T-2/2/c2c03b6165996e30312e5b7c7b681155

[13] P. Dreuw, P. Steingrube, H. Hanselmann, and H. Ney, "SURF-Face: face recognition under viewpoint consistency constraints," *British Machine Vision Conference*, Sep. 2009.

[14] V. Jain, E. Learned-Miller, and A. McCallum, "People-LDA: anchoring topics to people using face recognition," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1–8.

[15] T. Cour, B. Sapp, C. Jordan, and B. Taskar, "Learning from ambiguously labeled images," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 919–926.

[16] D. Ramanan, S. Baker, and S. Kakade, "Leveraging archival video for building face datasets," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1–8.

[17] D. Le and S. Satoh, "Unsupervised face annotation by mining the web," in *Data Mining, IEEE International Conference on*. Los Alamitos, CA, USA: IEEE Computer Society, 2008, pp. 383–392.

[18] W. Lin, R. Jin, and A. Hauptmann, "Web image retrieval re-ranking with relevance model," in *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*, 2003, pp. 242–248.

[19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[20] H. Bay, B. Fasel, and L. V. Gool, "Interactive museum guide," *Smart Environments and their Applications to Cultural Heritage*, p. 39, 2006.

[21] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, vol. 96, 1996, pp. 226–231.

[22] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," *SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, 2000. [Online]. Available: http://portal.acm.org/citation.cfm?id=335191.335388