Text Driven Recognition of Multiple Faces in Newspapers

Nicola Adami, Sergio Benini, and Riccardo Leonardi
Department of Information Engineering, Signals & Communications Lab
University of Brescia, via Branze 38, 25123, Brescia, Italy
Email: nicola.adami@ing.unibs.it, sergio.benini@ing.unibs.it, riccardo.leonardi@ing.unibs.it

***Abstract*—Face recognition is still a hard task when performed on newspaper images, since they often show faces in non-frontal poses, prohibitive lighting conditions, and too poor quality in terms of resolution. In these cases, combining textual information derived from the page articles with visual information proves to be advantageous for improving the recognition performance. In this work, we extract characters' names from articles and captions to restrict facial recognition to a limited set of candidates. To solve the difficulties derived from having multiple faces in the same image, we also propose a solution that enables a joint assignment of faces to characters' names. Extensive tests in both ideal and real scenarios confirm the soundness of the proposed approach.**

Keywords-Face recognition; Newspapers; Text analysis; Visual information; Multimodal.

I. INTRODUCTION

Over the last decades researchers have been making attempts to solve the problem of machine recognition of faces [1]. Algorithms proposed during the years can be coarsely classified into two categories: holistic approaches, such as Principal component analysis (PCA) [2] or Linear discriminant analysis (LDA) [3], and local feature-based ones (see e.g., [1] and [4]). Most recent developments of the holistic approaches include the Marginal Fisher analysis (MFA) [5], Eigenfeature regularization and extraction (ERE) [6], the sparse representation [7] and asymmetric PCA [8] and LDA [9], while Elastic bunch graph matching (EBGM) [10] and Active appearance model (AAM) [11] can be considered among the most performing feature-based algorithms. Despite the advances brought by these recent methods, there are still challenging problems to be tackled in face recognition, such as variations in pose, different facial expressions, make-up, lighting conditions as well as occlusions and cluttered background.

The recognition task is even harder when targeting newspaper images, where human faces are usually pictured at low quality and/or resolution. When visual data are not enough informative, one possible solution is combining natural language and visual information for improving semantic understanding of images. Newspapers in fact provide text that can be used to help the recognition process: each image with characters is commonly connected to an article on the same page, or at least described by a textual caption.

The idea has been relatively unexplored until the work in [12], which first proposes to use captions to locate faces in the accompanying photographs, thus with no recognition aims. A few years later, the rule-based PICTION system [13], trained on a dataset of 50 pictures was able to recognise human faces with a success rate of 65% by combining captions and photographs, even without employing a face recognition system.

In this paper, we target automatic recognition of human faces appearing in real newspapers by combining visual and textual information. As an advance with respect to previous work, we exploit all textual information coming from the newspaper page, thus not limiting the analysis to captions only, as done in the recent investigations in [14], [15], and [16]. Concerning face detection, we first apply an improved version of the Viola-Jones classifier [17] nowadays considered as a standard baseline for face detection. The recognition phase is then performed by using the standard method provided by Principal Component Analysis (PCA) [2]. These traditional approaches usually guarantee acceptable performance in not too complex contexts. However, these specific editorial products are mined by two major impediments that make harder the recognition process: one is related to the often too poor quality of images chosen for publishing in terms of resolution, contrast, illumination and pose; the second is due to the dimensions of the characters' database, which are potentially unlimited. These issues, which may lead to a difficult recognition, are here also addressed. Finally, as an additional contribution, recognition performance in case multiple faces are present in the same image are improved by a mechanism that jointly assigns identities to detected human faces.

Figure 1 describes the workflow: the textual analysis module extracts from newspaper pages potential characters' names to restrict the recognition phase to a few candidates. Results of recognition are inspectable by a human operator, who also takes care of the cases when a name (respectively, a face) found in the page has no associated face (resp. a name) stored in the databases, so that the system is able to learn from previous recognition processes.

Helping the automatic understanding of newspaper articles, such a tool could find application in complex tasks such as news segmentation, or in supporting professional frameworks for production of new multimedia content, such as news aggregators or feeds.

The document is organized as follows: in Section II, we

 ISBN: 978-1-61208-195-3

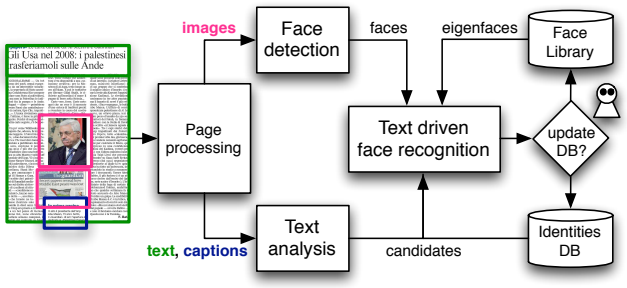

Figure 1: Diagram describing the workflow for the combined visual and text-based face recognition process.

explain how a segmented page looks like. Face detection and text analysis processes are described in Sections III and IV, respectively. In Section V, we focus on the multiple face recognition algorithm that employs a weak supervision in the form of text found in articles and captions. Experiments are conducted in Section VI, while conclusions are finally drawn in Section VII.

II. Page Preprocessing

Newspaper pages are first segmented by a process whose description is beyond the scope of this paper. As shown in Figure 2, the output of the segmentation stage consists of two separate pages (the *image-page* and the *text-page*) provided with two related *structure files* describing all page elements (*images*, *articles, titles*, *captions*, etc.) and their positions in the original page.

Figure 2: Example of how a) a newspaper page is segmented into b) an *image-page*, and c) a *text-page*.

III. Face Detection

In order to ensure a correct recognition of characters, a robust face detection process on the image-page becomes a fundamental prerequisite. Once segmented from the rest, the image-page is fed into a classifier to isolate regions belonging to human faces. This is achieved by relying on the well-known Viola-Jones method [17] endowed with the following adjustments to boost the detection performance.

On the one side we increase the robustness to pose change: despite the fact this classifier is known to work well for frontal faces only, the last implementation in [18] comes with several cascade files for detecting profile faces, even if with slightly lower performance. On the other hand, before being processed, images undergo a process of histogram equalization as in [19] to increase the algorithm invariance to complex lighting conditions. These adjustments are important in the specific domain of newspaper images, where characters are usually pictured in arbitrary poses, under different illumination conditions and often at low resolution.

Last but not least, since the whole system aims at recognising characters, each missed face during detection (i.e., each true negative) results in a final missed recognition. On the contrary, false positives are not that relevant, since they will not match with any candidate face in the database, therefore not producing errors. Thus the final tuning of the cascade face classifiers has been carried out in order to privilege *recall* rather than *precision*. Examples of *non-relevant* and *relevant* errors during the detection phase are given in Figure 3-a and Figure 3-b, respectively.

Figure 3: Examples of errors in detection: a) non-relevant error: the false positive on the tie will not find any matching face during recognition; b) relevant error: the missed detection of the face will determine a missed recognition.

IV. Text Analysis

Since in newspapers face recognition is a hard task, it is necessary to exploit all information we have at hand to support the identification of characters. Luckily, the text-page usually contains a direct reference to the characters' names pictured in the image-page. Even if this might not always happen, from our experiments (see Section VI) the case where a pictured character is not cited in the text is so rare that we can assume this hypothesis as reasonable.

To emulate the process of human comprehension while reading a newspaper is a hard task: a few decades of research dealing with the problem of natural language processing (see e.g., [20]) still have not solved the problem. For our purposes however, it is sufficient that the text module is able to extract names that might correspond to characters contained in images. To achieve this goal, the module relies on two databases: the *identity database* containing several characters' data, and one *common names database* [1].

[1]Lists of names and surnames are easily available for each country. In some languages their extraction is easier since they start with capital letters.

Whenever a name is found in the text, if that identity is already present in the characters' database, then that person is proposed as a *candidate* for the face recognition process. Conversely if a name is recognised as such but the related identity is not in the database, a new identity is proposed for approval to a human operator for insertion. In this case, a new database record is created and few images retrieved on the web (with Google's images search [21]) are used to populate the character's *face library*.

The identity database also manages name variations referring to the same person (e.g., "George W. Bush", "G. W. Bush", "George Bush", etc.), so that when any of these variants is found in the text, a unique *candidate* is selected for the recognition phase. When only the surname (e.g., "Bush") appears in the text, all characters with the same surname (e.g., the singer "Kate Bush") are chosen as candidates. Even if one surname is very popular, this will restrict anyway the candidates to a limited pool of characters.

In the common case when a caption accompanies the image, the importance of the text module increases, since it is sufficient to use the caption text as the only input (instead of the whole related article) to restrict the pool of candidates to a very short list, as shown in the example of Figure 4.

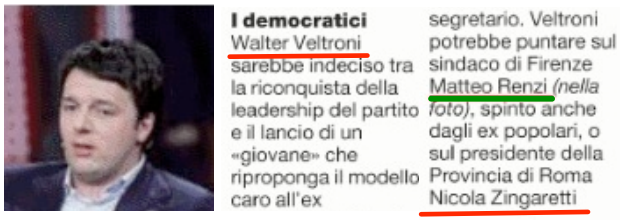

Figure 4: Caption returns a restricted pool of candidates: "Walter Veltroni", "Matteo Renzi", and "Nicola Zingaretti".

The module is not error free: for example it fails in case the text refers to "G. W. Bush" only as "President", or in general, in every case when a person is referred to by means of his/her title, such as "Prime Minister" or similar ones.

V. Text driven face recognition

Due to the fact that the application must be able to update the database during execution, the training phase on face recognition should be as automatic as possible. In fact, in case there is the need to add new faces, it is not reasonable to involve the human operator in too complex update operations. As a consequence, all local recognition approaches that require a manual labelling of points of interests (such as EBGM [10] or and AAM [11]) are not applicable in this context.

On the contrary, traditional holistic methods such as PCA [2] or LDA [3], if provided even with few image examples, are able to build sufficiently accurate models for each identity in an automatic way. Despite more modern approaches exist, performance offered by Principal Component Analysis in its original formulation are sufficient for implementing the visual part of the proposed supervised recognition approach.

A. Training

The initial face library contains a limited set of famous characters belonging to politics, sports, economy, science and culture. As seen in Section IV, each time that a new name pops up from newspaper pages, related face images retrieved on the Internet can be dragged in the face library. To ensure an acceptable level of automation, inserted pictures are automatically cropped and resized without forcing the user to extract faces manually: to do this the face detection algorithm described in Section III on equalized images first extracts face bounding boxes; after, image dimensions are normalized to 64×64 pixels. If the image quality is acceptable, faces are rotated and centered by using an eye detection algorithm, thus obtaining the final training image. Examples of processed images are shown in Figure 5.

Figure 5: 64×64 normalized faces are extracted for training.

Obtained images are then used for the training phase, which is performed by a standard PCA. Eigenfaces are calculated from the training set by keeping the M-images that correspond to the highest eigenvalues, so that they contain at least the 80% of the total energy, as suggested in [22]. These M eigenfaces define the M-dimensional "face space" employed during recognition. As new faces are added to the face database, eigenfaces can be updated or recalculated.

B. Recognition

Face recognition is treated as a pattern recognition task: each detected face Γ is projected onto the "face space" by transforming it into its eigenface components

$$\Gamma = [\gamma_1, \gamma_2, \ldots, \gamma_M]$$

which describe the contribution of each eigenface in representing the input face. The feature vector Γ is then used in a standard pattern recognition algorithm to find which of a number of predefined face classes best describes the face.

Since face recognition is a particularly difficult task, especially in case of numerous possible identities as in newspapers, to improve recognition performance we reduce

 ISBN: 978-1-61208-195-3

the number of face candidates only to those N identities found in the same text-page. Since it is rare that characters in pictures are not referred in the article body or in the caption itself, we accept the risk connected to an excessive candidate reduction.

The advantage of this supervised approach is most apparent when only one face is detected and one name is extracted from the text-page. As shown later in the experiments, this event is frequent when captions are associated to images: in such a situation, there is no need to perform a projection on the face space, but the image face is directly associated to the uniquely extracted name.

In case of multiple candidates instead, classification is performed by comparing the feature vector Γ of the test face with the face classes, which are the average representations

$$\overline{\Phi} = \left[\overline{\phi}_1, \overline{\phi}_2, \ldots, \overline{\phi}_N\right]$$

of each candidate over a number of face images. Comparison is based on the Mahalanobis' distance between Γ and $\overline{\Phi}$ to find which candidate best describes the test face:

$$k = \underset{i}{\operatorname{argmin}} \left\{ \sqrt{\left(\Gamma - \overline{\Phi}_i\right) S_i^{-1} \left(\Gamma - \overline{\Phi}_i\right)^T} \right\} i = 1, 2, \ldots, N$$

where S_i is the covariance matrix of each candidate class. By attributing more relevance to components with larger associated eigenvalues, this metric removes the problems related to scale and correlation that are inherent with the Euclidean distance, and provides in fact, superior experimental performance; in particular the average value is considered in order to better exploit each class distribution, and not relying only on a minimal distance that often leads to misclassification due to the presence of noisy samples. The value of the Mahalanobis' distance is also considered as a confidence level on the classification result: the user has the possibility to choose whether the confidence level is enough high for him not to check the recognition results, or conversely, if too low, to classify the face as *unknown* and add it manually to the face library for later use, so that the system learns to recognize new face images.

The same mechanism allows also for removing false positives introduced during face detection mentioned in Section III. By dividing the training space in two regions ("face" and "non-face" hemispaces) if the distance exceeds the space region boundaries, the detected face is probably a false positive, so that it is removed and recognition is not performed at all.

C. Joint recognition of multiple faces

The system as proposed so far is quite robust, especially until up to one face is detected in each image. In the presence of multiple faces in the same picture however, it is possible that two or more faces are at minimum distance to the same candidate. In this case, the algorithm as defined before, would label both faces with the same identity, thing that is evidently not possible. There is then the need to increase the algorithm robustness and elaborate a strategy to manage the recognition of multiple faces in the same image.

In order to best describe this problem, let us consider the example in Figure 6-a) where, due to his face orientation, Barack Obama is not recognised. In this picture, two faces are assigned to the same candidate "T. Geithner", since both are at a minimum average distance from Geithner's training samples (as shown in Figure 6-b).

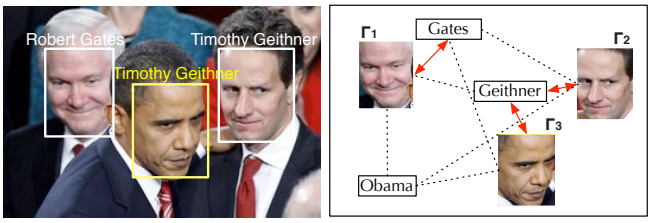

Figure 6: a) Independent recognition on three faces assigns the same identity to two characters, since b) both Γ_2 and Γ_3 are at minimum distance with the same candidate.

In order to increase the algorithm's robustness and correct results in analogous situations, we propose to consider also all other classes' distributions: to be assigned with an identity, a test image should not only be at minimum average distance from the candidate samples but also, at the same time, at maximum average distance from other classes.

This objective is achieved by assigning an heuristic score q to each candidate as the difference between the distance of the test image Γ and all other classes $\overline{\Phi}_i$ and the distance of Γ and the best candidate class $\overline{\Phi}_k$, that is

$$q(\Gamma) = \sum_{i=1}^{N} \left[d\left(\Gamma, \overline{\Phi}_i\right) - d\left(\Gamma, \overline{\Phi}_k\right)\right]$$

where score q is a real positive number. Once defined the heuristic score, the face recognition algorithm for multiple faces works as shown in Figure 7.

```
1. perform face recognition independently on all faces;
2. assign to each face the best candidate Φk;
3. If no conflicts,
4.     then exit;
5. else
6.     compute score q for each conflicting face Γ;
7.     assign face with highest q to identity Φk;
8.     remove Φk from the candidate list;
9.     repeat from line 2;
```

Figure 7: Joint face recognition algorithm.

To correct Obama's identity as in Figure 8-a, when both test faces Γ_2 and Γ_3 are at minimum distance with the same candidate "Geithner", the test face with higher q is

assigned to the best candidate, while the face with lower q is reassigned to the next closest candidate (Figure 8-b). Please notice that $q(\Gamma_2) > q(\Gamma_3)$ means that Γ_2 is on average further from other possible candidates than Γ_3, so that it is more likely that the latter was incorrectly assigned.

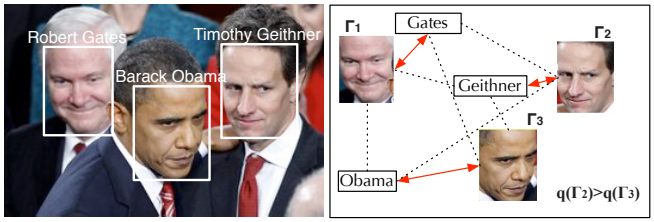

Figure 8: a) Obama's identity is corrected via joint face recognition; b) the face with higher q, i.e., Γ_2 is assigned to its best candidate, while Γ_3 is reassigned to the next candidate.

The algorithm is able to solve an arbitrary number of conflicts and works also when the number of faces is larger than the database possible identities. In this case, faces with worst q scores are labelled as *unknown*. Please notice that the label *unknown* does not necessarily imply that the identity is not in the database, but only that the face was not coupled with any of the candidates extracted from text.

VI. Experiments

The experimental part is subdivided in four progressive steps, so that to appreciate the beneficial brought by different modules of the supervised approach.

In the first part of the experiments, the performance of the face recognition algorithm are tested alone on a public available database "Olivetti-Att-ORL" [23], which contains 400 human faces belonging to 40 unique people, where each individual is pictured 10 times from a frontal view or with a slight tilt of the head (see Figure 9). Training

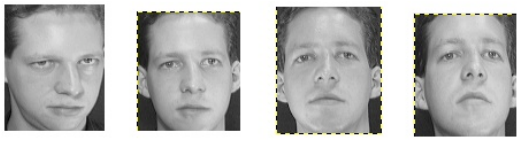

Figure 9: First test: recognition is performed on images from the Olivetti-Att-ORL dataset.

is performed on 325 randomly chosen images, while the recognition task is performed on the remaining part of the dataset. Classification results averaged on 5 runs returns 75% of correct recognition, which is in line with state of the art PCA performance [1].

In the second part of the experimental phase, we test the recognition algorithm on real newspaper pictures. For this aim a face library from newspaper images has been built as follow: around 200 identities have been extracted from different copies of the italian newspaper "Corriere della Sera" and for each identity, an average of four face images has been retrieved from the web, for a total training set of approximately 800 images. The test set instead is built up by using 200 images extracted from 15 different issues of the same newspaper, which contain faces in arbitrary conditions of pose, illumination, and quality, as shown in the examples of Figure 10. In this real application scenario, recognition performance collapse, as expected, around 50%, thus confirming that recognition of characters in newspaper images is far more challenging than on a standard dataset.

Figure 10: Recognition on newspaper images is more challenging due the variety of image poses, light conditions, or quality.

In the third step of the experiment, the approach combining face recognition with the supervision of text has been tested on the newspapers data. For each test image, we constrain recognition only to those candidates whose names are found in the same text-page or in the corresponding image caption. Results obtained yield a percentage of 83% correctly recognised faces.

Finally, the application to the same dataset of the algorithm for multiple face recognition further improves performance up to 86%. Table I summarises all performed tests and the related performance.

Table I: Performed experiments.

Test	Algorithm	Data	Perf.
1	*Face recognition only*	Olivetti-Att-ORL	75%
2	*Face recognition only*	Newspapers	50%
3	*Text driven recognition*	Newspapers	83%
4	*Joint supervised recog*	Newspapers	86%

The built system is able to learn from previous classifications by including the recognised faces in the face library. Therefore it is commonsensical to believe that performance are expected to improve as long as the system is in use.

As final considerations, we mention three causes of errors that influence the system performance. First, errors might be generated in case of wrong initial segmentation into text- and image-pages. For example, if one image is associated to a wrong caption, this likely leads to a missed recognition. Second, errors can be due to the face detection: true negatives, as seen in Section III are estimated to be around 10%. Third and last, the text module might fail in extracting correct names: this happened three times during the 200 performed tests (1.5%) because the mentioned identity was addressed only by his/her title. All these types of error are not included in results of Table I, which accounts only for the performance

 ISBN: 978-1-61208-195-3

of the supervised algorithm when both the text and the face detection modules return correct results.

VII. Conclusions

In this work, we combine text derived from newspaper articles and captions with visual information to improve face recognition performance. Characters' names are used to constrain facial recognition to a limited set of candidates, which are jointly assigned to the related faces in case multiple characters are present in the same picture. The good performance obtained in the experimental phase demonstrates that this approach allows for high recognition rate on newspaper images, notoriously a difficult benchmark since often showing faces in non-frontal poses, prohibitive lighting conditions, and poor in quality and/or resolution. Future work aims at extending the approach to a wider set of editorial publications (including magazines, satyrical, etc.) as well as to integrate higher performing recognition method.

References

[1] W.-Y. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

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

[3] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using LDA-based algorithms," *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 195–200, Jan. 2003.

[4] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognition*, vol. 39, no. 9, pp. 1725–1745, 2006.

[5] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.

[6] X. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 383–394, 2008.

[7] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[8] X. Jiang, "Asymmetric principal component and discriminant analyses for pattern classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 931–937, 2009.

[9] ——, "Linear subspace learning-based dimensionality reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16–26, 2011.

[10] L. Wiskott, J. M. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 775–779, 1997.

[11] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 23, no. 6, pp. 681–685, Jun 2001.

[12] V. Govindaraju, D. Sher, R. Srihari, and S. Srihari, "Locating human faces in newspaper photographs," in *Proc. of IEEE Conf. on CVPR*, San Diego, CA, Jun 1989, pp. 549–555.

[13] R. K. Srihari, "PICTION: A system that uses captions to label human faces in newspaper photographs," in *Press, A. (ed.) Proceedings of the AAAI-91*, 1991, pp. 80–95.

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

[15] T. Mensink and J. Verbeek, "Improving people search using query expansions: How friends help to find people," in *European Conference on Computer Vision*, ser. LNCS, vol. II. Springer, oct 2008, pp. 86–99.

[16] D. Ozkan and P. Duygulu, "Interesting faces: A graph-based approach for finding people in news," *Pattern Recogn.*, vol. 43, pp. 1717–1735, May 2010.

[17] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proceedings of CVPR*, vol. 1, pp. 511–518, 2001.

[18] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[19] L. Tao, L. M.-J. Seow, and V. K. Asari, "Nonlinear image enhancement to improve face detection in complex lighting environment," *International Journal of Computational Intelligence Research*, vol. 2, no. 4, pp. 327–336, 2006.

[20] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press, 1999.

[21] "Google's image search," http://images.google.com/. [retrieved: April, 2012]. Available: http://images.google.com/

[22] R. Tjahyadi, W. Liu, and S. Venkatesh, "Automatic parameter selection for eigenfaces," in *Proceedings of the 6th International Conference on Optimization: Techniques and Applications*, 2004.

[23] F. Ferdinando Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota, FL, Dec. 1994.

 ISBN: 978-1-61208-195-3