On the use of SIFT features for face authentication

Manuele Bicego, Andrea Lagorio, Enrico Grosso, Massimo Tistarelli
University of Sassari
{bicego,lagorio,grosso,tista}@uniss.it

Abstract

Several pattern recognition and classification techniques have been applied to the biometrics domain. Among them, an interesting technique is the Scale Invariant Feature Transform (SIFT), originally devised for object recognition. Even if SIFT features have emerged as a very powerful image descriptors, their employment in face analysis context has never been systematically investigated.

This paper investigates the application of the SIFT approach in the context of face authentication. In order to determine the real potential and applicability of the method, different matching schemes are proposed and tested using the BANCA database and protocol, showing promising results.

1. Introduction

Face recognition is possibly one of the first cognitive processes used by humans to recognize familiar people. Even though other sensorial cues are also adopted, like speech, gait and at birth even odor, the ability to recognize known faces is present already at birth. This and other issues make face recognition a very interesting an challenging research area in biometrics and computer vision.

Face recognition is certainly a complex problem, but essentially can be reduced to a pattern classification task. Many pattern recognition techniques have been applied and others have been developed ad hoc [22]. In the case of face analysis an additional complexity is due to several features of faces which are not common to other pattern recognition problems:

- the curse of dimensionality (at least one 2D image must be processed) is worsen by the variability of the pattern to be classified.
- The face is not a rigid object and it is continuously subject to non-rigid deformations.
- What makes faces different is also what they have in common, for example two eyes and a mouth.
- Even though a face is generally processed as a two-dimensional object, it is not, most ambiguities arise and some hypotheses fail because of the 3D structure of the face and its motion in space.

Due to these facts, the analysis of human faces is inherently an ill-posed problem [2]. For this reason, different techniques have been applied to constrain the pattern matching and classification processes. Among them, it is worth citing all methods based on the reduction of the face-space dimensionality by means of different optimization processes, such as the *Principal Component Analysis* (PDA), *Linear Discriminant Analysis* (LDA), *Fisher Discriminant Analysis* (FDA) and *Independent Component Analysis* (ICA) [17]. Other techniques are based on constraining and modeling the appearance of the face on the image, both as shape and texture information. Several methods have been based on the extraction and classification of salient facial features by means of multi-scale filtering with Gabor kernels [20, 4, 21, 9]. Along this direction, the techniques based on the estimation and progressive warping of a *"morphable face model"* explicitly derive a constrained mapping between the 3D face and its two-dimensional appearance on the image [16].

Recently, the *Scale Invariant Feature Transform* (SIFT) has emerged as a cut edge methodology in general object recognition as well as for other machine vision applications [13, 11, 12, 10, 5]. One of the interesting features of the SIFT approach is the capability to capture the main gray-level features of an object's view by means of local patterns extracted from a scale-space decomposition of the image. In this respect, the SIFT approach is similar to the *Local Binary Patterns* method [21, 9], with the difference of producing a view-invariant representation of the extracted 2D patterns.

Despite the wide applicability and potential of this technique, for the classification of 2D images, it was never applied to face recognition/authentication, at least to the best of our knowledge. In this paper a first attempt to apply the SIFT for face classification is reported. The basic SIFT scheme is tested on a standard face database [1] with three different matching techniques.

Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)
0-7695-2646-2/06 $20.00 © 2006 IEEE

In general, the a-priori knowledge of the geometry of the object to be recognize, can be successfully used to improve the recognition performances in both accuracy and speed [8, 15]. For this reason, the core SIFT algorithm has been adapted to the classification of face images according to three different schemes. In the proposed solutions the extracted features are selected and grouped according to the face geometry, driven by the knowledge of the position of few facial landmarks (typically the mouth and the eyes).

From the obtained results it is evident that the classification is more accurate when the information about the face geometry is used to drive the selection of the features. In this view, the real potential and applicability of this technique for face recognition is investigated.

2. Scale Invariant Feature Transform

In the 2004 David Lowe presented a method to extract distinctive invariant features from images [13]. He named them Scale Invariant Feature Transform (SIFT). This particular type of features are invariant to image scale and rotation, and are shown to provide robust matching across a a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination.

They are well localized in both the spatial and frequency domains, reducing the probability of disruption by occlusion, clutter, or noise. Large numbers of features can be extracted from typical images with efficient algorithms. A typical image of size 500x500 pixels will give rise to about 2000 stable features (although this number depends on both image content and choices of various parameters). In addition, the features are highly distinctive, which allows a single feature to be correctly matched with high probability against a large database of features, providing a basis for object and scene recognition. The cost of extracting these features is minimized by taking a cascade filtering approach, in which the more expensive operations are applied only at locations that pass an initial test. Following are the major stages of computation used to generate the set of image features:

1. **Scale-space extrema detection** The first stage of computation searches over all scales and image locations. It is implemented efficiently by using a difference-of-Gaussian function to identify potential interest points that are invariant to scale and orientation. Given a Gaussian-blurred image,

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$

where $I(x, y)$ is the given image and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{\sigma^2}}$$

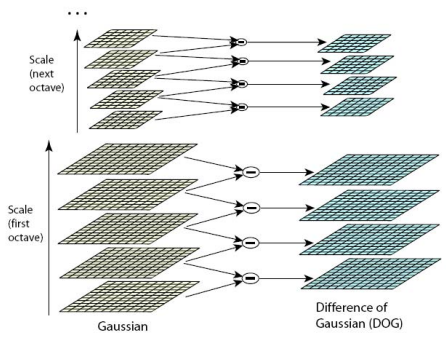


Figure 1. The blurred images at different scales, and the computation of the difference-of-Gaussian images (from [13]).

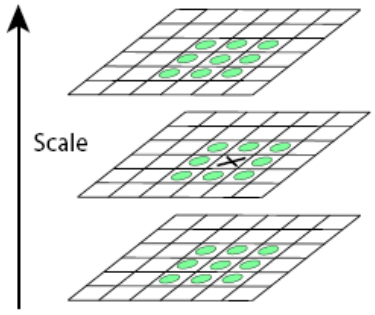


Figure 2. Local extrema detection, the pixel marked with a x is compared against its 26 neighbors in a 3 x 3 x 3 neighborhood that spans adjacent DoG images (from [13]).

to efficiently detect stable keypoint locations in scale space, the method proposed in [11] could be used. It makes use of the scale-space extrema in the difference-of-Gaussian function convolved with the image, D(x,y,), which can be computed from the difference of two nearby scales separated by a constant multiplicative factor k:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned}$$

Interest points (called keypoints in the SIFT framework) are identified as local maxima or minima of the DoG images across scales. Each pixel in the DoG images is compared to its 8 neighbors at the same scale, plus the 9 corresponding neighbors at neighboring scales. If the pixel is a local maximum or minimum, it is selected as a candidate keypoint.

2. **Accurate keypoint localization**: At each candidate location, a detailed model is fit to determine location and scale. Keypoints are selected based on measures of their stability.

   Once a keypoint candidate has been found by comparing a pixel to its neighbors, the next step is to perform a detailed fit to the nearby data for location, scale, and ratio of principal curvatures. This information allows points to be rejected when having low contrast (and therefore be sensitive to noise) or are poorly localized along an edge.

3. **Orientation assignment**: One or more orientations are assigned to each keypoint location based on local image gradient directions. To determine the keypoint orientation, a gradient orientation histogram is computed in the neighborhood of the keypoint (using the Gaussian image at the closest scale to the keypoint's scale). The contribution of each neighboring pixel is weighted by the gradient magnitude and a Gaussian window with a σ that is 1.5 times the scale of the keypoint. Peaks in the histogram correspond to dominant orientations. A separate keypoint is created for the direction corresponding to the histogram maximum, and any other direction within 80% of the maximum value. All the properties of the keypoint are measured relative to the keypoint orientation, this provides invariance to rotation.

4. **Keypoint descriptor**: The local image gradients are measured at the selected scale in the region around each keypoint. These are transformed into a representation that allows for significant levels of local shape distortion and change in illumination. Once a keypoint orientation has been selected, the feature descriptor is computed as a set of orientation histograms on 4×4 pixel neighborhoods. The orientation histograms are relative to the keypoint orientation, the orientation data comes from the Gaussian image closest in scale to the keypoint's scale. Just like before, the contribution of each pixel is weighted by the gradient magnitude, and by a Gaussian with σ 1.5 times the scale of the keypoint. Histograms contain 8 bins each, and each descriptor contains an array of 4 histograms around the keypoint. This leads to a SIFT feature vector with $4 \times 4 \times 8 = 128$ elements. This vector is normalized to enhance invariance to changes in illumination. In this way the descriptor is invariant to affine changes in illumination.

Some examples of the application of the SIFT algorithm to face images (database BANCA [1]) are shown in Fig. 3 and 4. In particular in the former three images of the same subject are displayed, showing that common features are present (see for example the black dots): there is a possibility of matching corresponding features. The latter images represent three different subjects: in this case SIFT features are very different.

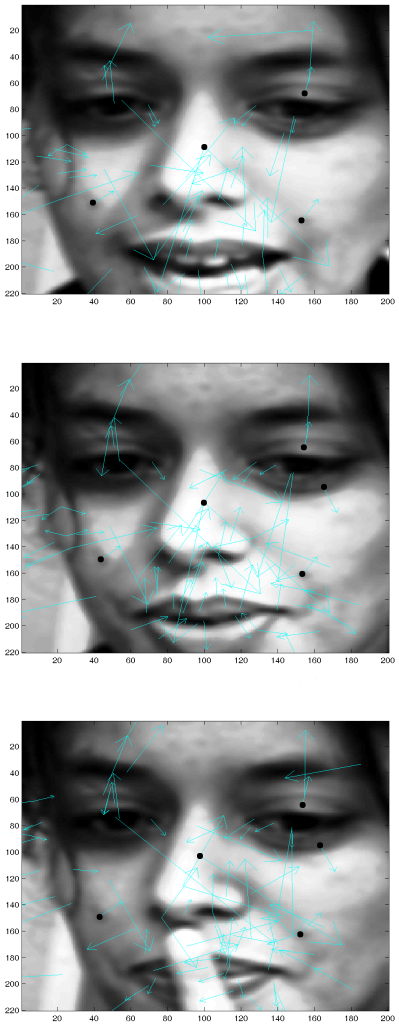


Figure 3. Example of images with extracted sift. The images represent the same subject in different pose. The black dot indicate some common stable SIFT for all three images

3. Matching strategies

To authenticate a face, the SIFT features computed in the test image should be matched with the SIFT features of the template. In this section different matching methodologies are investigated. They are different from the Lowe's method [13], in the sense that they are simpler and more related to

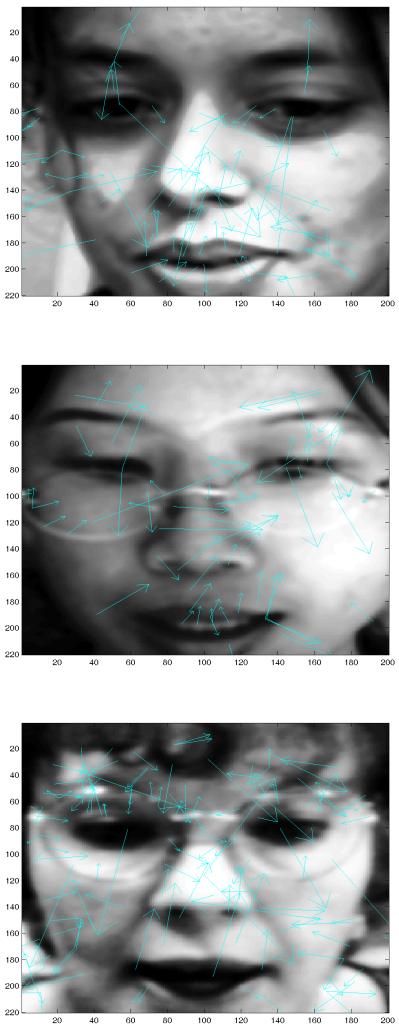


Figure 4. Example of images with extracted sift. The images represent different subjects. The SIFT are very different.

the problem we are addressing. Moreover the methodology proposed in [13] is devoted to recognition rather than authentication (recognition implies finding the best match, solved in [13] using a modified Hough Transform).

In each matching strategy the starting point is represented by two sets of features, computed on the testing and on the template images. As explained in Section 2, each feature is composed by four parts: the locus (location in which the feature has been found), the scale, the orientation and the descriptor. This last is a vector of 128 values. For simplicity, given a keypoint k_i, let's call $F(k_i)$, $L(k_i)$, $S(k_i)$ and $O(k_i)$ its feature descriptor, its location, its scale and its orientation, respectively.

The different methodologies could employ all or only a part of the whole information included in the SIFT feature.

3.1. Minimum pair distance

This methodology is the simplest one: computing the distance between all pairs of keypoint descriptors in the two images and use as matching score the minimum distance. More formally, given two images I_{test} and I_{temp}, representing the testing and the template images, respectively, two set of features are computed:

- $K(I_{test}) = \{k_1^{I_{test}}, k_2^{I_{test}} ... k_M^{I_{test}}\}$
- $K(I_{temp}) = \{k_1^{I_{temp}}, k_2^{I_{temp}} ... k_N^{I_{temp}}\}$

The matching score $D^{MPD}(I_{test}, I_{temp})$ (*Minimum Pair Distance*) is computed as

$$D^{MPD}(I_{test}, I_{temp}) = \min_{i,j}(d(F(k_i^{I_{test}}), F(k_j^{I_{temp}})))$$

where $d(F(k_i), F(k_j))$ is a distance between descriptors. In this paper the simple Euclidean Distance has been investigated, even if more complicated ones could be employed (for example Correlation — see [19]).

This simple scheme does not employ neither the location nor the scale and orientation information: it represents a real base-line system. The main idea under this method is that there is one point of the face which contains a very distinctive feature of the subject, which could be found in the testing image.

3.2. Matching eyes and mouth

This second method takes into account the fact that most part of the face information is located around the eyes and the mouth [3, 18, 14]. Once the position of these landmarks is determined, a matching strategy can be driven to consider only SIFT features belonging to such image areas, neglecting other less informative points.

Different techniques have been proposed for determining the position of eyes and mouth (see for example [6] and the references therein): here we assume that such positions are known. Given an image I, two sub images are extracted: one located around the eyes and one located around the mouth, called I^{eyes} and I^{mouth}, respectively. Then the matching is performed in a pair-wise manner, that is eyes with eyes and mouth with mouth. Finally the two distances are averaged. More formally:

$$\begin{aligned} D^{EM}(I_{test}, I_{temp}) &= \frac{1}{2} D^{MPD}(I_{test}^{eyes}, I_{temp}^{eyes}) + \\ &+ \frac{1}{2} D^{MPD}(I_{test}^{mouth}, I_{temp}^{mouth}) \end{aligned}$$

3.3. Matching on a regular grid

The first methodology does not take into consideration the location of the features: this represents a problem, since the two keypoints corresponding to the minimum distance could be not related to the same face part. In other words all parts of the face could be matched with all others, which is not realistic. This fact is alleviated in the second methodology, since only eyes and mouth are considered. Nevertheless also in this case features located on the right eye could be matched with features located on the left one. Therefore, if the images are more or less registered, a location-dependent matching could be performed. Registration is a particularly important problem in face authentication and recognition, and should be solved. Nevertheless it is completely a different problem from recognition / authentication, and should be solved before applying matching techniques. In fact, in all the recent databases (such as BANCA [1]) the positions of the eyes are given in order to permit the pre-registration of the images: only the matching methodology is analyzed. In this paper we assume already registered images.

The matching methodology presented in this paragraph subdivides the images in different sub-images, using a regular grid with overlapping. The matching between two images is then performed by computing distances between all pairs of corresponding sub-images, and finally averaging them. More formally, the two images are subdivided in a set of partially overlapped sub-images, called $I^1 ... I^T$. After a preliminary experimental evaluation (not shown here) we found that sub-images of dimensions 1/4 and 1/2 of width and height, respectively, represent a good compromise between accuracy of localization and possibility of recovering from registration errors. The overlapping was set to 25%.

Finally, the matching score $D^{RG}(I_{test}, I_{temp})$ (*Regular Grid*) is computed as the average between the matching scores computed on the pairs of images, namely:

$$D^{RG}(I_{test}, I_{temp}) = \frac{1}{T}\sum_{t=1}^{T}(D^{MPD}(I^t_{test}, I^t_{temp}))$$

4. Experimental evaluation

The following face authentication experiments were carried out on the BANCA database [1]—a multimodal database, containing both face and voice. The part used for face authentication is composed by 52 subjects (26 female and 26 male) For each subject, 12 different sessions were recorded under different conditions (4 *controlled*, 4 *degraded* and 4 *adverse*). For each session, 5 images were extracted, and used for training and for client and impostor testing.

In the BANCA protocol, 7 different experimental configurations have been defined, of increasing difficulty. In our experiment we used the Matched Controlled (MC) protocol, where the images gathered from the first session are used for training, whereas for testing images from second, third, and fourth sessions are employed. In this case we used registered images, that is images for which landmarks positions are known: this permit to register the images and concentrate only on recognition results. In particular, in the preprocessing phase, all the images were processed using a simple geometric normalization, followed by histogram equalization. In the geometric normalization, the face was mapped to the 210 pixel high by 200 pixel wide output image. The mapping used an affine transform—only translation, rotation and scaling. The image was transformed such that the manually annotated eye positions were mapped to points 25% in from the edges and 35% down from the top of the output image. The histogram equalization was carried out using a standard approach [7].

In order to get indicative results, the testing images are divided into two groups, G1 and G2, of 26 subjects each. The error rate was computed using the following procedure [1]:

- perform the experiment on G1, getting G1 scores
- perform the experiment on G2, getting G2 scores
- compute the ROC curve using G1 scores, determine the Prior Equal Error Rate and the corresponding threshold θ_{G1}
- use the threshold θ_{G1} to compute False Acceptance Rate ($\text{FAR}_{G2}(\theta_{G1})$) and False Rejection Rate ($\text{FRR}_{G2}(\theta_{G1})$) on the G2 scores
- compute the Weighted Error Rate (WER(R)) on G2 by determining

$$\text{WER}(R) = \frac{FRR_{G2}(\theta_{G1}) + R \cdot FAR_{G2}(\theta_{G1})}{1+R}$$

  for R= 0.1, 1 and 10
- compute WER(R) on G1 by a dual approach

The parameter R indicates the cost ratio between false acceptance and false rejection.

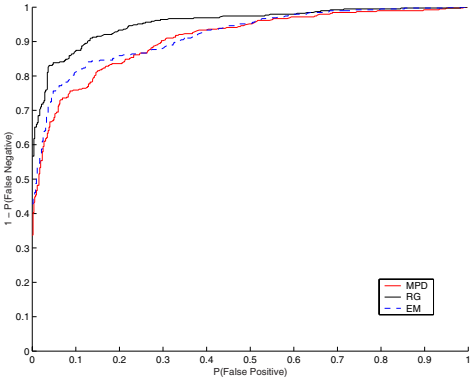
The SIFT features have been computed with Lowe's code[1]. Both the three matching methodologies have been tested: accuracies of authentication are proposed in Table 1 and 2. In particular, Prior Equal Error Rates for G1 and G2 are presented in Table 1 (the corresponding ROC curves are shown in Fig. 5), whereas Weighted Error Rates are proposed in Table 2, for three different values of R.

From the tables and the figures it is evident that taking into account the context information is beneficial: when

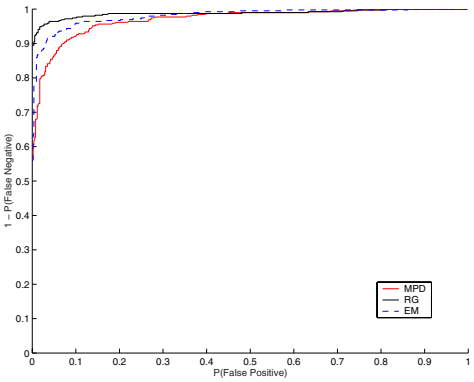
[1] Available at `http://www.cs.ubc.ca/ lowe/keypoints/`.

	MPD	EM	RG
Prior EER on G1	17.15%	15.38%	11.31%
Prior EER on G2	8.69%	6.38%	3.85%
Average	12.92%	10.88%	7.58%

Table 1. Prior EER on G1 and G2 for the three methods: 'MPD' stands for Minimum Pair Distance, 'EM' for Eyes and Mouth, 'RG' for Regular Grid.



(G1)



(G2)

Figure 5. ROC curves for G1 and G2: 'MPD' stands for Minimum Pair Distance, 'EM' for Eyes and Mouth, 'RG' for Regular Grid.

comparing corresponding parts of the face a significant improvement is obtained. In particular an improvement is gathered when concentrating the comparison only on mouth and eyes. Moreover, the best result is obtained with the Regular Grid methodology, that is when comparing corresponding parts. From these results emerges the crucial role played by the localization information in the matching.

	MPD	EM	RG
WER (R=0.1) on G1	22.56%	20.55%	15.97%
WER (R=0.1) on G2	7.40%	4.42%	3.04%
WER (R=1) on G1	17.95%	14.94%	10.51%
WER (R=1) on G2	9.52%	8.14%	6.35%
WER (R=10) on G1	13.33%	9.32%	5.06%
WER (R=10) on G2	11.64%	11.86%	9.65%

Table 2. Different WER for the three methodologies: 'MPD' stands for Minimum Pair Distance, 'EM' for Eyes and Mouth, 'RG' for Regular Grid.

5. Conclusions

In this paper the use of SIFT features in the context of face authentication has been investigated. Three different matching techniques have been proposed, namely:

- computing the distance between all pairs of keypoint descriptors in the two images and use as matching score the minimum distance.
- Use only SIFT features belonging to the areas around the eyes and mouth.
- The matching is performed considering the SIFT features located along a regular grid and matching overlapping patches.

The three techniques have been tested on the G1 and G2 image sets from the BANCA database. From the experiment carried out, the matching performed along a regular grid outperforms the other two methods, while the minimum pair distance gives the poorest results. Even though the obtained scores do not match the best face classifier tested on this database, still they confirm the applicability of the SIFT features in this context. It is worth noting that no accurate normalization of the illumination and shape has been performed.

From a first application of the SIFT features it appears that the feature matching process must be driven to cope for the peculiarities of the face shape and variability. On the other hand, the SIFT algorithm itself should be further analyzed and adapted to be fully tailored to the face shape and texture. This is a first attempt toward this direction, but more sophisticated matching techniques and the application of proper feature classifiers will be investigated in the future. In order to better understand the real potentiality of the method we will compare it with other methods (e.g. PCA or LDA). Another problem to be investigated in the future is the possibility of using SIFT to solve the registration problem.

References

[1] E. Bailly-Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Marithoz, J. Matas, K. Messer, V. Popovici, F. Pore, B. Ruiz, and J.-P. Thiran. The BANCA database and evaluation protocol. In *Proc. Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA03)*, pages 625–638. Springer-Verlag, 2003. 1, 3, 5

[2] M. Bertero, T. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76(8):869–889, Aug. 1988. 1

[3] I. Biederman. Human image understanding: Recent research and theory. *CVGIP*, 32:29–73, 1985. 4

[4] J. Bigun. Retinal vision applied to facial features detection and face authentication. *Pattern Recognition Letters*, 23(4):463–475, 1997. 1

[5] M. Brown and D. Lowe. Recognising panoramas. In *IEEE Int. Conf. on Computer Vision*, pages 1218–1225, 2003. 1

[6] P. Campadelli and R. Lanzarotti. Fiducial point localization in color images of face foregrounds. *Image and Vision Computing*, 22:863–872, 2004. 4

[7] R. Gonzalez and R. Woods. *Digital Image Processing*. Prentice Hall, 2 edition, 2002. 5

[8] W. Grimson. *Object Recognition By Computer-The Role Of Geometric Constraints*. Mit Press, Cambridge, 1990. 2

[9] G. Heusch, Y. Rodriguez, and S. Marcel. Local binary patterns as an image preprocessing for face authentication. IDIAP-RR 76, IDIAP, 2005. 1

[10] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2004. 1

[11] D. Lowe. Object recognition from local scale-invariant features. In *Int. Conf. on Computer Vision*, pages 1150–1157, 1999. 1, 2

[12] D. Lowe. Local feature view clustering for 3d object recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 682–688, 2001. 1

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004. 1, 2, 3, 4

[14] F. Nahm, A. Perret, D. Amaral, and T. Albright. How do monkeys look at faces? *Journal of Cognitive Neuroscience*, 9:611–623, 1997. 4

[15] V. Nalwa. *A Guided Tour of Computer Vision*. Addison Wesley, 1993. 2

[16] S. Romdhani, V. Blanz, C. Basso, and T. Vetter. Morphable models of faces. In S. Li and A. Jain, editors, *Handbook of Face Recognition*, pages 217–246. Springer Verlag, 2004. 1

[17] G. Shakhnarovich and B. Moghaddam. Face recognition in subspaces. In S. Li and A. Jain, editors, *Handbook of Face Recognition*, pages 141–168. Springer Verlag, 2004. 1

[18] M. Tistarelli. Active/space-variant object recognition. *Image and Vision Computing*, 13(3):215–226, 1995. 4

[19] M. Tistarelli, A. Lagorio, and E. Grosso. Understanding iconic image-based face biometrics. In *Biometric Authentication*, volume LNCS 2359, pages 19–29. Springer Verlag, 2002. 4

[20] L. Wiskott, J. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-19:775–779, 1997. 1

[21] G. Zhang, X. Huang, S. Li, Y. Wang, and X. Wu. Boosting local binary pattern (lbp)-based face recognition. In L. 3338, editor, *SINOBIOMETRICS 2004*, pages 179–186. Springer Verlag, 2004. 1

[22] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35:399 – 458, 2003. 1