

Building Statistical Appearance Models Using Residual Information

Andreas Lanitis*

Department of Computer Science and Engineering, Cyprus College, Cyprus
Email: alanitis@cycollege.ac.cy

Abstract. Statistical appearance models are generated by applying statistical analysis on the color and shape variation in an ensemble of image objects belonging to the same class. Statistical appearance models have proved useful for locating, reconstructing and interpreting image objects that undergo systematic appearance variation. In this paper we demonstrate how residual information can be incorporated in the model building process in an attempt to generate models robust to local occlusions and models that can deal effectively with subtle types of variation. With our approach we first train a statistical appearance model using the standard technique and then estimate the residual obtained by reconstructing the samples from the training set. The analysis of the residual information allows us to define the areas of the object that the model fails to model correctly. Further statistical analysis is applied locally to such areas of the object. Quantitative results prove that statistical appearance models built using the proposed method are capable of reconstructing more accurately the appearance of previously unseen objects belonging to the same class.

1 Introduction

Statistical appearance models [6, 8, 9, 11] are generated by applying statistical analysis on the color and shape variation of an ensemble of image objects belonging to the same class. The process of generating such models involves the application of Principal Component Analysis in order to define the most important modes of color and shape variation among the training set. Once a model of this form is trained, it can be used as the basis for compact and reversible representation of image objects, hence the problem of processing/representing such objects is reduced to the problem of processing their model-based representation.

Statistical appearance models based on this form have been used for modeling appearance variation of a wide range of variable image objects such as human faces [2, 7, 9, 12, 18] and other biological organs [11]. Once models of this form are generated they can be used in numerous applications including the detection and location of variable image objects in unseen images [6, 7], interpretation of image objects [12, 18] and reconstruction and synthesis of the appearance of objects [2, 13, 14].

* Research supported by the Research Promotion Foundation (Cyprus) – project 26/2001

Statistical appearance models based on Principal Component Analysis (PCA) suffer from various disadvantages the most important of those being the inability to model correctly local, but in several cases important, types of variation. During the analysis emphasis is paid to the areas of the object that display the most dramatic appearance variation, where as less intensive types of appearance variations are suppressed. Also PCA based models do not deal effectively with local non-systematic deformations or occlusions of image objects. Since the transformation of the shape and color of an object to the corresponding model parameters is done by considering the appearance of the whole object, local unpredicted variations and/or occlusions affect the overall coding process. As a result PCA based appearance models are not able to reconstruct accurately the appearance of locally deformed and/or occluded image objects. Another disadvantage of statistical appearance models is the computational time and memory requirements for processing matrices of large size (at least in the case that objects are modeled in high resolution). As a result the use of high-resolution statistical appearance models in real time applications is inhibited.

In this paper we propose a variation to the basic method of generating statistical appearance models in an attempt to generate models that deal effectively with local occlusions and subtle appearance variation. With our approach we first train a statistical appearance model using the standard technique and then estimate the residual obtained by reconstructing the samples from the training set. We use residual information in order to define the areas of the object that the model fails to model correctly. According to the intensity of the residual in different regions, the area of the object is segmented to different parts. Further statistical analysis is applied to the segmented areas of the object in an attempt to generate models that deal effectively with local appearance variation and occlusion. The main steps involved in the proposed methodology are illustrated in Figure 1. Quantitative results prove that by incorporating residual information during the process of generating statistical appearance models it is possible to produce models capable of reconstructing more faithfully the appearance of previously unseen objects belonging to the same class.

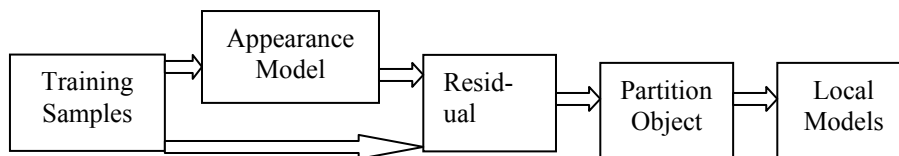


Fig. 1. Overview of our approach

Since the work presented is related to statistical appearance models, in the remainder of the paper we first present the standard method for generating such models and then present a brief literature review related to the subject. The modification of the basic approach is described in section 4 and experimental results are presented in section 5. Conclusions related with our work and our plans for future work in this area are described in section 6.

2 Statistical Appearance Models

Cootes et al [4, 6, 8, 9] describe how Statistical Appearance Models are generated by applying Principal Component Analysis on the shape and intensity of a set of training examples. The basic method of building PCA-based models and the way in which models of this type form the basis of Statistical Appearance Models are briefly presented in the following paragraphs. A detailed description of the procedure of building and using combined Statistical Appearance Models is presented elsewhere [4, 6, 8, 9].

2.1 PCA-based Statistical Models

Statistical models are generated from a set of training examples represented by a set of variables. We calculate the mean example (\mathbf{X}_m) among the training set and establish the deviation of each example from the mean. Principal component analysis (PCA) is applied on the covariance matrix of the deviations so that the main ways in which training examples vary from the mean example are extracted. As a result of the analysis training examples can be reconstructed using:

$$\mathbf{X} = \mathbf{X}_m + \mathbf{P}\mathbf{b} \quad (1)$$

Where \mathbf{X} is a training example, \mathbf{X}_m is the mean example, \mathbf{P} is the matrix of eigenvectors and \mathbf{b} is a vector of weights, or model parameters. By solving equation 1 with respect to \mathbf{b} , it is possible to calculate the set of model-parameters corresponding to a given example. This approach can be used for compact and reversible parameterization of training examples.

2.2 Modeling Shape and Intensity Variation



Fig. 2. Example of landmarks used for training a shape model



Fig. 3. Example of shape-free patches used for training an intensity model

In order to train a shape model, a number of landmarks are placed at each training sample at predefined locations. For example locations of landmarks used for training a face shape model is shown in Figure 2. In this case the vector \mathbf{X} (i.e the representation of each training sample) contains the x and y co-ordinates of the landmarks. Before the application of the process indicated in section 2.1, all training examples are aligned in order to minimize shape variation due to rotation, scaling and translation.

The alignment is done by finding the best affine transformation coefficients for each training shape so that the point-to-point distance between training shapes is minimized. The alignment process is described in detail in [4]. Once a shape model is trained the shape of similar objects can be parameterized using the shape parameters, i.e the parameters of the shape model.

During the process of training an intensity model, all training images are warped to the mean shape among the training set, so that shape variation within the training set is eliminated. The intensities within a pre-specified image region are used for training an intensity model. Figure 3 shows examples of shape-normalized shape-free face patches, used for training a face intensity model. In order to build a grey-scale intensity model the vector X contains the grey-level intensities at each pixel within the shape-normalized object area, whereas in the case of color models the vector X contains the red, green and blue components of each pixel [9]. It is worth mentioning that intensity models of this type are similar to the models reported and used by Turk and Pentland [18].

2.3 Generating an Appearance Model

Unlike the models described earlier that model shape or intensity separately, an Appearance model models the overall appearance of objects. In order to build a Statistical Appearance Model we train individual shape and intensity models and approximate all training examples to the corresponding model parameters of the shape and intensity models. As a result of this representation the X vectors used for training an Appearance Model contain the shape and intensity parameters of the corresponding training samples. Based on this representation it is possible to train an integrated Appearance Model that models both the shape and intensity of objects [6].

3 Literature Review

Recently many researchers reported variations to the basic method of building and using statistical appearance models, in an attempt to make them applicable to a wider range of applications. Efforts in improving those models can be classified into two main streams: the ones that aim to use PCA based analysis locally, rather than applying the method globally and the ones that aim to build models using other image features instead of using just the shape and intensity values of the objects to be modeled. A brief review of related work is presented in the following sections.

A number of researchers propose the application of the PCA decomposition locally rather than globally. In this content local spaces are defined either by spatial segmentation of an object or by dividing the training set into different clusters and applying PCA individually on different clusters. In the case of spatial segmentation, the area of the objects to be modeled is segmented into different parts representing local features of interest and different models are trained for each part [15, 16]. Recently Avidan [1] proposed a new method for segmenting the area of the object where instead of partitioning the object area into predefined regions, the segmentation is

done based on the variation in the intensity of each pixel. In his work the separation of the pixels into different segments is done based on the correlation in the variation of pixel intensities among the training set. As a result pixels that display similar behavior over the training set are grouped together in order to form a segment and PCA analysis is then applied locally to each segment. It is important to note that pixels belonging to the same segment are not necessarily adjacent pixels. Experimental results demonstrate that when models based on this methodology are used, it is possible to obtain more accurate reconstructions and better classification results. Since our method bears similarities to the work reported by Avidan [1] further discussion and comparison of the two methods is provided in section 6.

Statistical models of this type are based on a linear formulation; hence they are applicable only in the cases that the variation in the training set is linear. If this is not the case the model will fail to model faithfully the appearance variation of the objects in question. A typical example of such cases is an attempt of modeling extreme 3D pose variation of human faces since the shape deformations involved in that case are highly non linear. A number of researchers proposed techniques based on the partition of the training set in different clusters and the statistical analysis of each cluster separately [3, 7]. By splitting the training set into clusters containing the most similar objects the variability in the local clusters can be assumed to be linear.

Statistical models are usually trained based on the grey level-intensities of the object to be modeled [6] or in the case of dealing with color the variables are the RGB components of each pixel [9]. Several researchers investigated the use of alternative representations of training samples, in order to generate models that represent more faithfully features of interest. Cootes and Taylor [5] describe statistical appearance models built using the orientation of the edge at each object pixel, in an attempt to generate models that emphasize features located on strong edges rather than features located on areas of uniform intensity. They demonstrate that models based on this approach can be used for locating structures in images more accurately, rather than conventional models. Stegmann and Larsen [17] represent the objects using the so-called VHE representation. In this context the V and S components are based on the intensity and saturation component in the HSV space. The E component is the edge strength of each pixel. Experimental results prove that this representation outperforms the standard AAM technique in locating objects in images.

4 Generating Models Using Residual Information

The ability of a model to reconstruct faithfully the appearance of objects from the training set, gives as an indication of the performance of a model. In our work we make use of the reconstruction errors as part of a self-correcting approach used for improving the performance of appearance models. In this context we first use the training samples to train an appearance model representing the overall object area. Based on the reconstruction errors obtained among the training set we define the areas of the object that it is difficult to model and the object area is partitioned to different segments, based on the residual information. We then train local models for each segment, thus we end up with a bank of local models instead of a single global model.

Further analysis is performed in order to train a new model that incorporates information among all local models. The main steps involved in the process are shown in Figure 1 and a detailed description is provided in the following sections.

4.1 Using Residual Information for Partitioning the Object Area

Once we train an appearance model we code all training samples into their model-based representation and subsequently reconstruct the appearance of the each sample in the training set. The residual in intensity between the original and reconstructed examples is given by calculating the difference in the intensity between the original sample and the reconstructed sample using equation 2.

$$R_i = | X_i - X_i' | \quad (2)$$

where X_i is a vector containing the intensity of all pixels of the i^{th} training sample and X_i' is the reconstruction of the intensity of the i^{th} training example. The R_i vector contains the absolute reconstruction error for each pixel.

Typical examples of intensity residual images obtained are shown in Figure 4. The bright areas in the images in Figure 4 show the areas of high residual, i.e the areas displaying variation that the model cannot cope with properly.



Fig. 4. Typical examples of residual images

Equation 3, is used for calculating the average residual pattern (R_m) among the k samples in the training set.

$$R_m = \frac{1}{k} \sum_{i=1}^k R_i \quad (3)$$

The average residual pattern provides an indication of the ability of the model to model correctly image areas among the whole training set. Example of the mean residual patterns for different training sets of face images are shown in Figure 5. Usually areas of high residual are located in areas of high spatial frequencies i.e near strong edges or in areas that it is common to encounter occlusions (i.e near the hairline when dealing with face images).

Based on the required number of segments to be used and the minimum and maximum residual values in the mean residual image, we establish the range of residual values corresponding to each segment and classify all pixels of the mean residual image to the appropriate segments. Figure 6 shows the allocation of pixels in each segment for different number of segments. (The darker areas correspond to areas of low residual whereas the brighter areas are the areas of high residual).

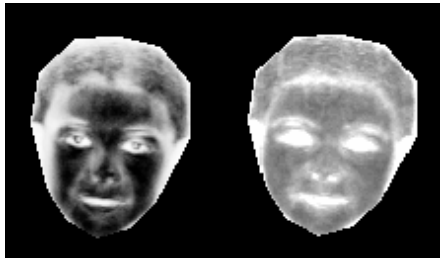


Fig. 5. Examples showing the mean residual pattern for two training sets of face images

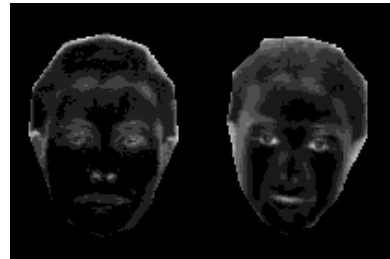


Fig. 6. Examples showing the segmentation of the object area to different segments (6 and 10 segments for the images in the left and right)

4.2 Training a Composite Appearance Model

Based on the residual based segmentation of the object area we segment the shape-free intensity image of each sample from the training set, to the corresponding segments and train a local intensity model for each segment. Those local models should be able to model more accurately the appearance of the corresponding object segments, since the models now are specific to a given segment.

Based on this representation, each example from the training set is represented by a set of shape parameters and a set of intensity model parameters for each segment. Following the example of Edwards et al [6, 8] we aim to train a single model so that the final representation we obtain will refer to a single model that incorporates information from the shape and the local intensity models. By training a single model we also ensure that we model correctly variation in the cases that the shape and intensity variation is correlated and also in the case that intensity variations between different segments is correlated. In order to train a global model we first represent all training samples in terms of the shape and the model parameters of each segment intensity model and we concatenate those parameters in a single vector. Before we form the concatenated vector we normalize parameters resulting from different model, in order to ensure that all models receive equal significance. By applying PCA decomposition on the resulting representation we generate a single model that models both the shape of the objects in the training set and it also models the intensity variation in each segment of the training objects.

4.3 Coding/Reconstructing Image Objects

In order to obtain the model-based representation of a new object, similar to the ones in the training set, the following procedure takes place.

- I. Based on the locations of the landmarks (assuming that the locations are already available) calculate the shape model parameters.

- II. Warp the object to the mean shape and extract the intensity values within the shape-free area.
 - III. Based on the residual based segmentation (see section 4.1), the shape free intensities corresponding to each segment are collected.
 - IV. The intensity model parameters for each segment are calculated.
 - V. Shape and intensity segment parameters are weighted and concatenated in a single vector.
 - VI. Based on the combined model the final model parameters are calculated.
- The process of reconstructing the appearance of an object, given a set of model parameters is as follows:
- I. Based on the combined model, the vector containing the shape and segment parameters is calculated.
 - II. Shape parameters and intensity parameters corresponding to each segment are extracted from the vector obtained in step I. Those parameters are used for estimating the shape of the object and the shape-free intensities of each segment. During the process a reverse-normalization step takes place in order to compensate for the normalization of the parameters taking place during the training procedure.
 - III. The intensities of each segment are merged in order to generate the overall shape-free object appearance.
 - IV. The overall shape-free appearance is warped to the shape defined in step II, in order to get the full reconstruction.

5 Experimental Evaluation

We have performed experiments in order to assess the performance of Statistical Appearance Models generated based on the framework outlined in the previous section. As part of our experiments we generate appearance models using both the standard method reported by Cootes et al [6] and the proposed method and test the accuracy of each method in reconstructing the appearance of training and unseen objects. In the following sections we describe the datasets used for our experiments, describe the experimental process and present the results.

5.1 Datasets Used in our Experiments

During the experimental evaluation we used datasets containing face images. Face images are complex structures undergoing different kinds of appearance variation, including occlusions due to spectacles and facial hair. Another feature encountered in datasets containing face images is the significant difference in appearance variation in different parts of the face. For example the variation in hairstyles is enormous when compared to the variation in the internal face. In our experiments we have used the following datasets:

Dataset 1: Dataset 1 [10] contains 100 color face images in total. In most cases there is only a single image per person, although there are about five persons that have more than one image in the dataset. The images in this dataset were collected under carefully controlled conditions so that lighting conditions, image resolution, background and facial pose are uniform in all images. The database contains subjects of both genders, subjects of different ethnic origins with ages ranging between 18 to approximately 50 years old. Although in some cases there are faces with beards, moustaches and/or spectacles, most faces are not occluded. Typical examples from dataset 1 are shown in Figure 7.

Dataset 2: Dataset 2 contains 400 color face images in total. The database contains multiple images of approximately 30 subjects. Images in this database were captured under varying lighting conditions, variable background, variable facial pose and variable resolution. The database contains subjects of both genders with ages ranging between 1 to 35 years old. Typical examples from dataset 2 are shown in Figure 7.



Fig. 7. Typical images from dataset 1 (top row) and dataset 2 (bottom row)

We have chosen to use two datasets for our experiments in order to assess the performance of the proposed method under different conditions.

5.2 Experimental Procedure

We evaluated the performance of the proposed methodology by assessing the ability of models generated to reconstruct the appearance of training objects and unseen objects. During the experimental evaluation each dataset was split into two subsets and each part was used for training a model while the second part was used for testing the ability of the model to reconstruct the appearance of previously unseen objects. The images from the datasets were divided in such a way so that images belonging to the same subjects appear only in one of the two database parts. Examples of reconstructions obtained are shown in Figure 8.



Fig. 8. Original images reconstructions from dataset 1 (top row) and dataset 2 (bottom row) and the corresponding reconstructions. The reconstructions on the left side are reconstructions of images from the training set whereas the reconstructions on the right side are reconstructions of previously unseen face images.

Each time the appearance of an object is reconstructed we calculate the overall reconstruction error by considering the image intensities within the object area, between the original image and the reconstructed image, as shown in Figure 9. The object area is defined as the area enclosed by the points of the shape model (see Figure 2) in the original and reconstructed images.

The performance of a model is defined by calculating the average reconstruction error among the training and the test set. In order to perform the experiments on a larger number of samples the training and test sets for each database were reversed so that for each database we run two experiments – the result quoted show the mean reconstruction error among both cases.



Fig. 9 . Original image (left) and the reconstruction (middle) . The reconstruction error is calculated by calculating the differences in intensity between the reconstructed and original image, within the area of the object (right image).

5.3 Experimental Results

Dataset		1	2
Number of Cases		100	400
Standard Method (Cootes et al [6])	Error (Training set)	25.6	25.9
	Error (Test set)	37.5	30.9
Proposed Method	Number of Segments	6	12
	Error (Training set)	23.4	22.2
	Error (Test set)	34.2	26.3

Table 1. The results of our experiments.

Table 1 shows the average reconstruction errors obtained for each method. In both cases models based on local intensity models defined based on residual information, outperform the models generated using the conventional method. In both cases there are optimum numbers of segments for which the models achieve the best performance. It is interesting to note that according to our experimental evaluation the optimum number of segments for each case minimizes the reconstruction error both in the training and the test set.

Although we get remarkable decrease in the reconstruction error when we use the proposed method, the number of parameters required to represent an object remains almost constant when compared with the number of parameters required to represent an object using models built based on the conventional method.

The computational cost for coding and reconstructing the appearance of objects using the proposed method is considerably lower than in the case of modeling the overall intensity variation using a single model. In the proposed approach the size of matrices that they need to be processed is significantly lower, resulting in faster processing. The exact computational gain we obtain depends on the number of segments and the distribution of pixels in each residual segment. In our experiments the computational gain is about 20% lower, when compared with the computational load of the standard method.

6 Conclusions

We have proposed a new method for building statistical appearance models. During the model building process we make use of the residual information in order to define the areas of the object that the model cannot cope correctly with. As a result the object area is segmented to a number of parts where each part represents the object pixels that the model can model with approximately equal accuracy. By training local intensity models for different segments we end up with models that they are specific to a certain object area, achieving in that way better explanation of the intensity variability in the training set. Experimental results demonstrate that models built based on this methodology can be used for reconstructing more accurately the appearance of objects from the training set and also the appearance of unseen objects. In other words such models display improved specificity to the training set and at the same

time they display better generalization abilities – both of these features are desirable in Statistical Appearance Models. The number of segments required depends on the training set. In the case of dataset 1 that contains image captured under carefully controlled way a smaller number of segments is required, whereas for noisy data (i.e dataset 2) more segments are required. Usually there is correspondence between the segments and the physical characteristics of the objects (see Figure 6) depending on the difficulty to model adequately different object features. Although our method was tested only on face images, it is generic and it can be used when modeling other types of variable objects.

Our approach bears similarities with work reported recently by Avidan[1]. Along similar lines he describes how the area of the objects to be modeled is split into segments and he demonstrated that when training multiple local segment models instead of a single global intensity model, the resulting models are better. The main differences between Avidan's approach and our method, is the process of defining the segments among the object area. Avidan segments the area of an object based on the correlation of intensities among different pixels, so that pixels displaying similar variation in the training set are grouped together. In our approach we segment the object area based on residual information. Although Avidan's approach is reasonable we believe that our methodology can be advantageous, since in our work the segmentation process is directly linked to the ability of the model to model correctly different object areas, thus we deal with the problem in a direct way. Also residual information provided naturally a self-correcting measure rather than having to define explicitly other rules for segmenting the object area. Based on the framework reported by Avidan an object in the class is represented by the parameters of the local models. In our work we train a global model based on the representation of the local segments, capitalizing in this way on possible intensity correlations between segments. Also in our case the model incorporates both shape and intensity variation, unlike the work reported by Avidan, where he deals only with intensity variation.

In the future we plan to perform more work in the area in order to deal with several issues that could lead to the generation of improved models. In particular we plan to apply a similar technique during the process of training the shape model, so that different subsets of points are used for training local shape models rather than using a global shape model. Also we plan to use optimization methods in order to establish the optimum number of segments to be used and the corresponding range of residuals in each segment. During the process we plan to take into account the variance of the residual at each pixel rather than using the average residual value only. We also plan to run extended evaluation tests in order to assess the performance of the system both against the standard method for training statistical appearance models and also against methods reported recently [1, 5, 17].

Acknowledgements

Our research in the area is supported by the project PROSOPO (Project 26/2001) funded by the Research Promotion Foundation of Cyprus. Dataset 1 contains images from the Psychological Image Collection at Stirling [10].

References

1. S. Avidan. "EigenSegments: a Spatio- Temporal Decomposition of an Ensemble of Images". *Proceedings 7th European Conference on Computer Vision*, Springer LNCS, Vol 2352, (2002) 747-758
2. V. Blanz, T. Vetter. A Morphable Model for the Synthesis of 3D Faces. *Proceedings Conference on Computer Graphics*, (1999) 187-194
3. R. Cappelli, D. Maio, D. Maltoni. "Multi-Space KL for Pattern Representation and Classification". *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 23(9):977-996, (2001)
4. T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham. Active Shape Models - their Training and Application. *Computer Vision Graphics and Image Understanding*, 61(1):38-59, (1995)
5. T.F. Cootes, C.J. Taylor. On Representing Edge Structure for Model Matching. *Proceedings IEEE Computer Vision and Pattern Recognition Conference*, (2001) 1114-1119
6. T.F. Cootes, G.J. Edwards, C.J. Taylor. "Active Appearance Models". *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 23:681-685, (2001)
7. T.F. Cootes, G.V. Wheeler, K.N. Walker, C.J. Taylor, "View Based Appearance Models". *Image and Vision Computing*, 20(9):658-664, (2002)
8. G.J. Edwards, A. Lanitis, C.J. Taylor, T.F.Cootes. Statistical Face Models: Improving Specificity. *Image and Vision Computing*, 16(3):203-211, (1998)
9. G.J. Edwards, T.F.Cootes, C.J.Taylor. "Advances in Active Appearance Models". *Proceedings 7th International Conference of Computer Vision*, Vol.I, Kerkyra, Greece, (1999) 137-142
10. <http://pics.psych.stir.ac.uk/>. Psychological Image Collection at Stirling, (2003)
11. <http://www.imm.dtu.dk/~aam/>. Active Appearance Models, (2003)
12. A. Lanitis, C.J. Taylor, T.F. Cootes, "Automatic Identification and Coding of Human Faces Using Flexible Models". *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 19(7):743-756, (1997)
13. A. Lanitis, C.J. Taylor, T.F. Cootes. "Toward Automatic Simulation of Aging Effects on Face Images". *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 24(4): 442-455, (2002)
14. A. Lanitis. "PROSOPO – a Face Image Synthesis System". In: Y. Manolopoulos, S. Evripidou, A. Kakas (eds.), *Advances in Informatics*. Post-proceedings 8th Panhellenic Conference in Informatics. Springer LNCS Vol.2563. (2003)
15. A.M. Martinez. "Recognizing Imprecisely Localized, Partially Occluded, and Expression Variant Faces from a Single Sample per Class". *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 24(6):748-763, (2002)
16. A. Pentland, B. Moghaddam, T. Starner. "View-Based and Modular Eigenspaces for Face Recognition". *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, (1994)
17. M.B. Stegmann, R. Larsen. "Multi-band Modeling of Appearance". *Image and Vision Computing*, 21:61-67, (2003)
18. M. Turk, A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3 (1):71-86, (1991)