Age Regression Based on Local Image Features

Azza MOKADEM
azza@telecom-paristech.fr

Maurice Charbit
maurice.charbit@telecom-paristech.fr

Gérard Chollet
gerard.chollet@telecom-paristech.fr

LTCI, UMR514, GET-Télécom
75014, Paris - France

Abstract

Human age estimation using facial image is becoming more and more investigated because of its potential applications in many areas such as multimedia communication and human computer interaction. Since many factors contribute to the aging process like gender, race, health, living style, the current age estimation performance for computer vision systems is still not efficient enough for practical use. In this paper, we addressed the problem of age estimation from single facial gray-scale image since the color information appeared as not significant in considered low resolution images. Local and global Discrete Cosinus Transformation (DCT) are used for feature extraction allowing thus a first dimensionnality reduction through this discriminative representation.A second reduction of dimensionality has been obtained through principal component analysis (PCA). A linear regression function has been learned and tested on different large databases extracted from MORPH [16]. Experimental results have shown some encouraging results.

Index Terms: age estimation, facial image features, PCA, regression

1. Introduction

Given a facial image, humans have the ability to accurately recognize and interpret information in real time. Multiple attributes can be estimated from it such as gender, expression, ethnic origin and age. These facial attributes are important since they play a crucial role in real applications including multimedia communications and Human Computer Interaction. For example, considering the application of age estimation, an Age Specific Human Computer Interaction system can be developed. This system has useful applications such as internet security when preventing kids to access adult pages. A vending machine can also refuse to sell alcohol or cigarettes to underage people. However, automatic age estimation remains a challenging task. The difficulty is in the way different people age since the aging process is determined not only by the person's genes but could be affected by external factors such as health, living style, living location, weather conditions and artifacts like makeup and plastic surgery. This paper uses classical approach for extracting significant facial attributes and performing age estimation on parameters obtained by projection into a low dimensional space.

The rest of the paper is organized as follows: first, the related work is briefly reviewed in section 2. The section 3 deals with the data representation. Section 4 introduces the method used to reduce data dimension and the regression methods. Section 5 presents the database used in this work and the experimental results. Finally, in section 6, conclusions are drawn.

2. Related work

The existing methods on the age estimation using face images can be divided into three categories: anthropometric model, aging pattern subspace and age regression.

The anthropometric approach [12] is based on the facial development theory and facial skin wrinkle analysis to classify images. The results obtained in [12] are 81.6% of accuracy for classification in three groups and only 27% of accuracy for classification in five groups. We noticed that the previous methods were tested on small databases (200 images).

In the second approach, aging pattern as defined in [6], is a sequence of personal images sorted in time order. That means that all images come from the same person and are ordered by time. Estimation of age is then performed by positioning the image, under testing, on all aging patterns and selecting the one with the lowest reconstruction error. This method also works with incomplete aging patterns. Two algorithms are derived, AGES (for aging pattern sub-

space) and AGES_{lda} depending or not on Linear Discriminant Analysis (LDA) [8], [6], [7]. In [5], the best result in term of Mean Absolute Error (MAE), which is the average absolute difference between the real age and the estimated one, is 5.36 years obtained on the FG-NET database (1002 images) [1].

The last approach we refer as age regression, consists of extracting facial features by projection in a discriminative subspace [14], [13]. Then, an image is represented by a set of parameters. Various regression functions are applied on these extracted features to estimate the age.

One of the best results [4] reaches an MAE of about 6 years using Conformal Embedding Analysis (CEA) associated with the quadratic regression function. Unfortunately, this result has been tested on a private database of 2000 images. Let us also refer to two similar age regression methods WAS (Weighted Age Specific Architecture) [13] and AAS (Appearance Age Specific Architecture) [13] where the faces were modelled by a set of shape and intensity parameters.

In addition to regression methods, authors also considered some conventional classification methods including k-Nearest Neighbors [17], Back Propagation neural network [21], C4.5 decision tree [19], and support vector machine [24]. In [6], all these methods were trained on the FG-NET database and tested on $1,724$ images extracted from the $55,608$ images of MORPH database but authors did not indicate which ones.

let us mention particularly the recent large survey [3] where authors present complete state-of-the-art techniques in the face-image-based age synthesis and estimation topics.

In our work, we only considered age regression approach where training and test were both performed on $5,000$ images extracted from the MORPH database (see section 5.1).

3. Data Representation

While in the previous works, images were just represented by their gray levels, resized with lower dimension and often cropped before projection into descriminative subspace, we decided in this work to extract at first features from images before projecting them.

Therefore, three facial representations have been used. The first one is the brute facial images in grayscaling levels. Because it is also known that Discrete Cosine Transform (DCT) is well suited for facial processing [18], [23], both global and local facial representations based on DCT were performed. In the global representation, entire image has been represented by the most significant coefficients of the DCT whereas in the local one, see also spatially flexible patch (SFP), in [23], the image was divided in small blocks and the most significant coefficients were selected to represent each block. In the three cases, the face feature components are reshaped as a p column vector x also called image in the following.

4. Dimensionality reduction and Regression

In this work, we considered facial images of resolution at most 65×60. In this context, authors [4], [9] [10], usually consider gray-level images.

In the following learning and testing examples are denoted $\{x_k, \ell_k\}$ where k is an index, x_k the image vector and ℓ_k the corresponding age. To prevent overfitting problems, dimensionality reduction has been applied. At first, we have considered Partial Least Squares (PLS) approach [15] because in PLS, the dimensionality reduction basically takes into account the correlation between the explicative variables and the response. Nevertheless, our obtained results with PLS have been not significantly different from those obtained with Principal Component Analysis (PCA) where dimensionality reduction only deals with explicative variables. Hence, PCA has been retained for reduction. Using this technique, the p-dimensional vector x is transformed into a d-dimensional vector y with $d < p$.

4.1. Principal Components Analysis

In data processing, PCA is a projection-based tool classically used to reduce dimensionality. The PCA principle is to find a d-rank projector that maximizes the dispersion of the projected data. That writes

$$\Pi^\star = \arg\max_{\Pi \in \mathcal{P}_d} \sum_{k=1}^{n} (x_k - \bar{x}) \Pi (x_k - \bar{x})^T \qquad (1)$$

where $\mathcal{P}_d$ denotes the set of d-rank projector on $\mathbb{R}^p$ and where $\bar{x} = n^{-1} \sum_{k=1}^{n} x_k$. The solution of this problem is given by $\Pi^\star = UU^T$ where $U = \begin{bmatrix} u_1 & \cdots & u_d \end{bmatrix}$ is the set of d eigenvectors associated to the d largest eigenvalues of the empirical covariance matrix $R = n^{-1} \sum_{k=1}^{n} (x_k - \bar{x})(x_k - \bar{x})^T$. Once the various directions and the size of learning images projection subspace determined, training and testing images were projected on it, allowing thus a dimensionality reduction.

4.2. Regression functions

Linear regression

Based on the reduced image $y = U^T x$, the age ℓ is assumed to fit the following linear model with linear dependence on y:

$$\ell = \beta_0 + \beta_1^T y + \epsilon \qquad (2)$$

where $\beta_0 \in \mathbb{R}$ is a scalar, $\beta_1 \in \mathbb{R}^d$ and where ϵ denotes the model noise which is assumed to be centered. Let us denote $\hat{\beta}_0$ and $\hat{\beta}_1$ the values estimated on the training database using least square criterion and $\hat{\ell} = \hat{\beta}_0 + \hat{\beta}_1^T y$ the predicted

age value associated to the reduced image y. The age estimation performance is evaluated by the Mean Absolute Error (MAE) defined by:

$$\text{MAE} = \frac{1}{n}\sum_{k=1}^{n}\left|\ell_k - \hat{\ell}_k\right| \tag{3}$$

Quadratic regression

We also considered, as in [4], linear regression model but with quadratic dependence on y

$$\ell = \beta_0 + \beta_1^T y + \beta_2^T (y \odot y) + \epsilon \tag{4}$$

where $\odot$ is the Kronecker product, $\beta_0 \in \mathbb{R}$, $\beta_1 \in \mathbb{R}^d$ and $\beta_2 \in \mathbb{R}^d$. Again β_0, β_1 and β_2 were estimated on the training database using least square criterion. Then the predicted age is given by $\hat{\ell} = \hat{\beta}_0 + \hat{\beta}_1^T y + \hat{\beta}_2^T (y \odot y)$.

4.2.1 Regression Tree

The regression tree [11] yields a M-partition of the explicative variables space $\mathbb{R}^d$ in M regions denoted R_1, R_2, ..., R_M corresponding to the M terminal tree nodes, such that the predicted age of an image y is given by

$$\hat{\ell} = \sum_{m=1}^{M} c_m \mathbb{1}_{R_m}(y) \tag{5}$$

where $\mathbb{1}_A$ is the indicator function of set A and

$$c_m = \frac{\sum_{k=1}^{n} \ell_k \mathbb{1}_{R_m}(y_k)}{\sum_{k=1}^{n} \mathbb{1}_{R_m}(y_k)} \tag{6}$$

At each step, the algorithm performs the binary split of the region R associated to some node in two sub-regions

$$\Lambda_1(j,s) = \{y \in R : y_j > s\} \text{ and } \Lambda_2(j,s) = R - \Lambda_1(j,s) \tag{7}$$

where y_j denotes the j-th component of y and $s \in \mathbb{R}$. Then the splitting variable j, the splitting threshold s and the values c_1 and c_2 are obtained by minimization of some impurity criterion as the following least squares criterion [11]

$$\sum_{\{k:y_k \in \Lambda_1(j,s)\}} (\ell_k - c_1)^2 + \sum_{\{k:y_k \in \Lambda_2(j,s)\}} (\ell_k - c_2)^2 \tag{8}$$

This partitioning is applied to each internal node. The process continues until each node contains at most a user-specified number of examples. In our work this number has been obtained by cross validation. We notice that regression tree was not yet applied in the field of age estimation.

4.2.2 AdaBoost

AdaBoost is basically used for classification problem. It is an algorithm for constructing a strong classifier as linear combination of simple weak classifiers. At each step, a weak classifier minimizing the training error on the weighted examples is chosen and the weight distribution is recalculated by increasing the weights of misclassified examples and by decreasing the weights of well-classified examples. Process stops when any new classifier performs worse than the pure chance. The Adaboost paradigm has been also considered for regression problem. In [22] the authors present two algorithms denoted AdaBoost.R2 and AdaBoost.RT. In AdaBoost.R2, all weights are updated at each iteration while in AdaBoost.RT the weights are updated according to some threshold, in such a way to emphasis on the only examples hard to classify. Moreover, the loss function of Adaboost.RT is computed using relative error rather than absolute error, as it is in AdaBoost.R2, making possible to give enough attention to the examples with low weights. Despite the fact that AdaBoost.RT is more complex because we need to calibrate the threshold, its main advantage is that its implementation is easily derived from that of the AdaBoost binary classifier. In our work the weak learner is obtained via a PCA by varying the number of principal components. The number of possible weak learners is 1000 and the final learner consists of 100 weak learners. We notice also that these two methods of adaboost for regression were not yet used for age estimation.

5. Experiments

5.1. Data Base

The database used in this work is the MORPH database [20]. This is a public database containing $55,608$ face images of different individuals with resolution of 240×200. These individuals are men and women of different races (Black, White, Hispanic, Indian, Asian or other), wearing or not glasses, with or without facial hear. The ages are ranging from 16 to 74 years. To avoid large computational time consumption, we used only some subsets of the MORPH database consisting of 5000 images with lower resolution. We also considered different homogeneity factors as it follows:

database DB1: 36×36 heterogeneous images with different gender and different races,

database DB2: 65×60 homogeneous images by selecting individuals which are male, black, without glasses and without facial hair,

database DB3: 36×36 homogeneous images by selecting only individuals which are male, black, without glasses and without facial hair.

These images are chosen so that the associated age vector is uniformly distributed through the different age classes. For training and testing, respectively $4,000$ and $1,000$ images are used and 5-fold cross-validation is applied. The images of the homogeneous and heterogeneous database used to train and test our algorithm are available in [2]

5.2. Age estimation experiments

In this paper we have retained the six following experiments:

Experiment 1: we used database DB1 and divided each image into 6×6 non-overlapping blocks. Each block was represented by the five largest DCT coefficients. That gives $5 \times 6 \times 6 = 180$ DCT coefficients.

Experiment 2: we used database DB1 and divided each image into 6×6 non-overlapping blocks. Each block was represented by the 30 largest DCT coefficients. That gives $30 \times 6 \times 6 = 1,080$ DCT coefficients.

Experiment 3: we used database DB3 and divided each image into 6×6 non-overlapping blocks. Each block was represented by the 30 largest DCT coefficients. That gives $30 \times 6 \times 6 = 1,080$ DCT coefficients.

Experiment 4: we used database DB2 and represented each image by all gray-scale pixels.

Experiment 5: we used database DB2 and represented each image by 1500 DCT coefficients.

Experiment 6: we used database DB2 and divided each image into 5×5 non-overlapping blocks. Each block is represented by $25 \times 13 \times 12 = 3,900$ DCT coefficients.

In figures 1 to 6, associated respectively to experiments 1 to 6, we have reported the MAE as a function of the reduced dimension when applying the linear regression. The plain curve shows the variation of MAE with the number of principal components on the training database whereas the dashed curve shows the variation of MAE with the number of principal components evaluating on the test database.

Table 1 shows the results of applying the regression tree. Table 2 shows the results of applying AdaBoost.RT and Adaboost.R2. Table 3 summarizes our best result and some others results from literature [6].

5.3. Discussion

According to the figures 1 and 2, we note that the greater the number of DCT coefficients is, the more we improve the results. Indeed, for images divided into 6×6 non-overlapping blocks, the MAE decreases from 9.76 years to 9.15 years when we keep from five to 30 largest DCT coefficients. As expected, the more we extract information on the local regions of the image, the best results of discrimination we achieve.

Figure 3 we have reported the MAE as a function of the number of PCA components in the same condition as in figure 2, except that we used now a homogeneous population. As expected, the homogeneity of the population makes the estimation more accurate. Indeed we observe that the MAE decreases from 9.15 years (figure 2) to 8.40 years (figure 3).

Figures 4 and 5, it appears that the DCT on the global image does not improve the performance compared to the brute image. We see that the MAE raises from 8.9 years for images represented by all their pixels (figure 4) to a MAE of 10.30 years for these same images represented by the 1500 largest DCT coefficients (figure 5).

From the figures 5 and 6, we can conclude that local characteristics are much more discriminative than the global ones, since the MAE decreases significantly from 10.3 years (figure 5) to a MAE of 7.61 years for images represented by all the DCT coefficients of 5×5 blocks (figure 6).

Tables 1 and 2 we have reported the MAE for the six experiments using respectively regression trees and both AdaBoost.R2 and AdaBoost.RT. We notice that all results are very close to the results using PCA with linear regression, whatever the factors involved. Concerning AdaBoost approach, these results are disappointing if we compare to the well-known efficiency of AdaBoost classifier. To this aim, we suggest to investigate other features, such as Haar features, successfully used in face detection.

Finally in table 3, we have reported our best results obtained on homogeneous database with 65×60 pixel images, divided into 5×5 blocks, each one represented with 25 DCT coefficients. For comparison we have also reported different published results [6] also tested on the MORPH database. Even if the protocol used in [6] is not completely specified, the results may be compared: our approach brings a slight amelioration, a MAE of 7.61 years, compared to the best result of AGES_{lda} method with a MAE of 8.07 years [6].

6. Conclusion

In this paper, we have proposed a new framework for automatic age estimation based on facial images attributes. A new data representation associated with PCA as a method for learning low dimensional age manifold was introduced. The linear regression was investigated for age prediction based on the learned manifold. From the experimental results, we conclude that our method is slightly better than the current state-of-the-art.

References

[1] FG-NET Aging Database. http://sting.cycollege.ac.cy/~alanitis/fgnetaging/index.htm. 2

Exp.1	Exp.2	Exp.3	Exp.4	Exp.5	Exp.6
10.14	10.10	9.8	9.59	9.9	7.92

Table 1. Mean Absolute Error in years of the age estimation on the MORPH database using Regression Tree for the six experiments (see section 5.2).

	Exp.1	Exp.2	Exp.3	Exp.4	Exp.5	Exp.6
A.R2	10.73	10.91	9.97	9.7	12.14	9.17
A.RT	10.75	10.1	9.14	9.91	11.73	8.25

Table 2. Mean Absolute Error in years of the age estimation on the MORPH database using AdaBoost.RT (A.RT) and AdaBoost.R2 (A.R2) for the six experiments (see section 5.2).

Method	AGES	AGES$_{lda}$	WAS	AAS	kNN
	8.83	8.07	9.32	20.93	11.3
Method	BP	C4.5	SVM	Our [2]	
	13.84	12.69	9.23	7.61	

Table 3. Mean Absolute Error in years of the age estimation on the MORPH database. Comparison with other methods (see [6]).

[2] protocol used. http://arbolifmd.webuda.com/. 4, 5

[3] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, January 2010. 2

[4] Y. Fu and T. Huang. Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia, vol.10, no. 4*, pages 578–584, 2008. 2, 3

[5] X. Geng and K. Smith-Miles. Facial age estimation by multilinear subspace analysis. *ICASSP*, pages 865–868, 2009. 2

[6] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *ACM Conf. Multimedia (ACM MM'06)*, pages 2234–2240, 2007. 1, 2, 4, 5

[7] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Facial age estimation by non linear aging pattern subspace. *Proceeding of the 16th ACM international conference on Multimedia*, pages 721–724, 2008. 2

[8] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. *ACM Conf. Multimedia (ACM MM'06)*, pages 307–316, 2006. 2

[9] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 17, NO. 7*, pages 1178–1188, July 2008. 2

[10] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang. A probabilistic fusion approach to human age prediction. *Computer Vision and Pattern Recognition Workshops. CVPRW '08.*, pages 1–6, July 2008. 2

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of statistical learning*. Springer, 2009. 3

[12] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. *Computer Vis. Image Understand., vol. 74, no. 1*, pages 1–21, 1999. 1

[13] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Trans. Syst, Man, Cybern.B., vol. 34, no. 1*, pages 621–628, 2004. 2

[14] A. Lanitis, C. Taylor, and T. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Trans. Pattern Anal. Machine Intell., vol. 24, no. 4*, pages 442–455, 2002. 2

[15] S. Maitra and J. Yan. Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *CAS Spring Meeting*, 2008. 2

[16] MORPH. http://www.faceaginggroup.com/projects-morph.html. 1

[17] E. Patrick, F.P.Fisher, and C.Christodoulou. A generalized k-nearest neighbor rule. *Proc. Information and Control, vol. 16, no. 2*, pages 128–152, 1970. 2

[18] V. Perrier. http://www.ljk.imag.fr/membres/Valerie.Perrier/SiteWeb/node9.html. 2

[19] J. Quinlan. C4.5. *Proc. Programs for Machines Learning*, 1993. 2

[20] K. Ricanek and T. Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. *Proc. Seventh Int'l Conf. Automatic Face and Gesture Recognition*, pages 341–345, 2006. 3

[21] D. Rumelhart, G.E.Hinton, and R.J.Williams. Learning representations by backpropagating errors. *Proc. Nature, vol. 323, no. 9*, pages 318–362, 1986. 2

[22] D. L. Shrestha and D. P. Solomatine. Experiments with adaboost.RT, an improved boosting scheme for regression. *Neural computation, vo .18*, pages 1678–1710, 2006. 3

[23] Y. Shuicheng, L. Ming, and T. Huang. Extracting age information from local spatially flexible patches. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.*, pages 737–740, 2008. 2

[24] V. Vapnick. *Proc. Statistical Learning Theory*. J. Wiley and Sons, 1998. 2

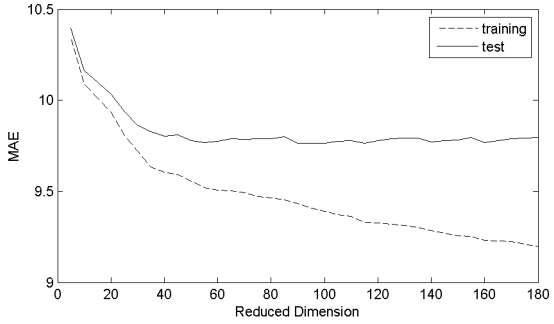

Figure 1. Experiment 1: 36×36 heterogeneous images with local transformations. The image is divided into 6×6 non over-lapping blocks. Each block consists of the 5 largest DCT coefficients. MAE is reported as a function of the PCA component number.

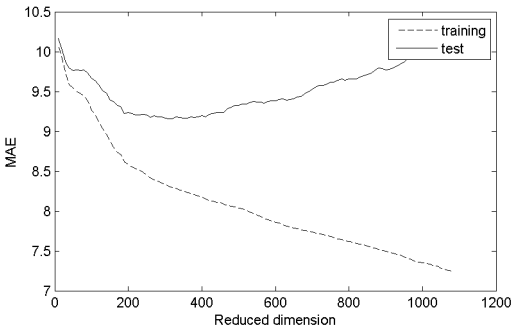

Figure 2. Experiment 2: 36×36 heterogeneous images with local transformations. The image is divided into 6×6 non over-lapping blocks. Each block consists of the 30 largest DCT coefficients. MAE is reported as a function of the PCA component number.

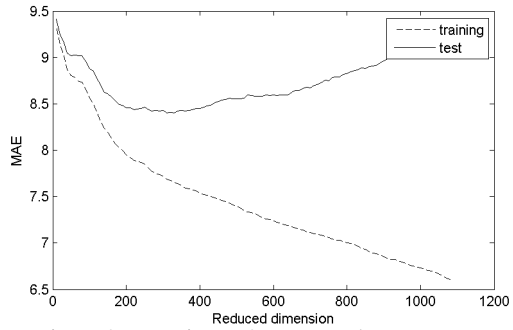

Figure 3. Experiment 3: 36×36 homogeneous images with local transformations. The image is divided into 6×6 non over-lapping blocks. Each block consists of the 30 largest DCT coefficients. MAE is reported as a function of the PCA component number.

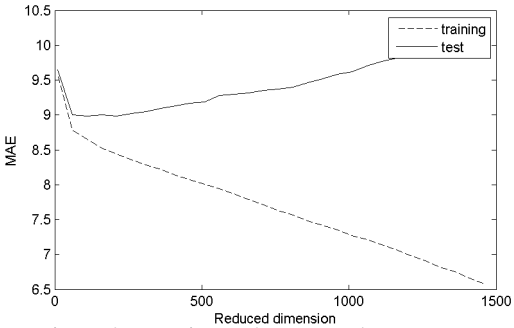

Figure 4. Experiment 4: 65×60 homogeneous images with gray levels. The image is represented by $3,900$ gray-scale pixels. MAE is reported as a function of the PCA component number.

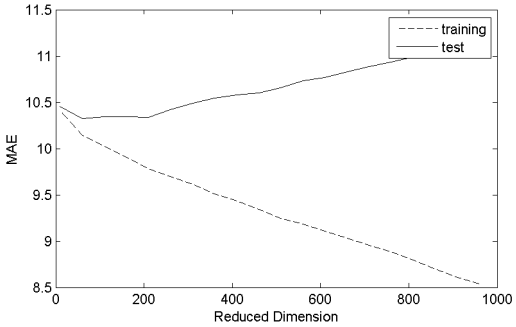

Figure 5. Experiment 5: 65×60 homogeneous images with global transformations. The features are $1,500$ DCT coefficients. MAE is reported as a function of the PCA component number.

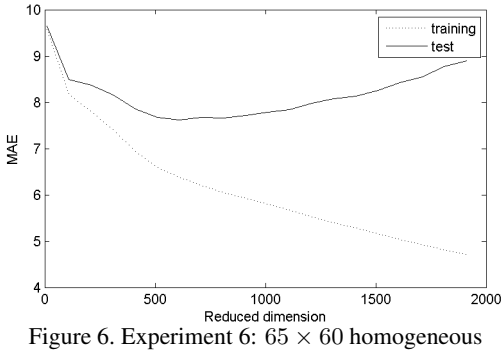

Figure 6. Experiment 6: 65×60 homogeneous images with local transformations. The image is divided into 5×5 non over-lapping blocks. The features are $25 \times 13 \times 12$ DCT coefficients. MAE is reported as a function of the PCA component number.