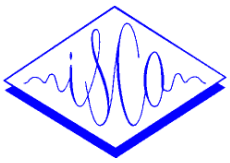

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Linking the structure and perception of 3D faces: Gender, ethnicity, and expressive posture

*Guillaume Vignali, Harold Hill, Eric Vatikiotis Bateson**

ATR HIS Laboratories, Kyoto, Japan
*Also UBC Linguistics Dept., Vancouver, Canada
xvignali@atr.co.jp, hill@atr.co.jp, bateson@atr.co.jp

Abstract

This paper reports a statistical study of human face shape whose overall goal is to identify and characterize salient components of facial structure for human perception and communicative behavior. A large database of 3D faces has been constructed and is being analyzed for differences in ethnicity, sex, and posture. For each of more than 300 faces varying in race/ethnicity (Japanese vs. Caucasian) and sex, nine postures (smiling, producing vowels, etc.) were recorded. Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) were used to reduce the dimensionality of the data and to provide simple, yet reliable reconstruction of any face from components that correspond to the sex, ethnicity, and posture of the face. Thus, it appears that any face can be reconstructed from a small set of linear and intuitively salient components. Psychophysical tests confirmed that the shape is sufficient to estimate sex and ethnicity. Subjects were asked to judge the sex and ethnicity of (a) natural faces and (b) faces synthesized by randomly combining Principal Component coefficients within the database. Subjects successfully discriminated ethnicity and sex independently of posture, verifying that different combinations of components are required and in differing amounts. Finally, implications of these results on animation and face recognition are discussed, incorporating results of studies currently underway that examine the "face print" residue of the sex-ethnicity factor analysis.

1. Database Presentation

The face of a speaker conveys many kinds of information about both the speaker and the content of what is being said. Much of our work at ATR has focused on the contribution of time-varying events on the face – e.g., the behavior of the lips, jaw, cheeks, and head motion [1]. Other studies have indicated the importance of other time-varying landmarks such as the eyes and the eyebrows in providing contextual and paralinguistic cues (e.g., [2]). As yet, however, we know very little about what role more global structural features such as the 3D shape of the face play in visual communication, whether shape is considered statically or as it varies over time. Although we ultimately want to examine how face shape influences intelligibility of the spoken signal, in this paper we apply Principal Components Analysis (PCA) on our 3D database of static faces to determine the space of variation and Linear Discriminant Analysis to make categorizations within that space. We also report perceptual experiments confirming the psychological validity of the artificially computed space and of the category specific dimensions within it.

The database consists at present of more than 310 3D laser scans of static heads. We used a Cyberware scanner that acquires 24bit texture maps and 3D polygon meshes of the head with about 25 000 points per head. Each subject is scanned for nine specific facial postures. Each scan takes about 17 seconds, so only postures that can be maintained for long periods can be used. Even with this limitation, the resulting set of postures covers a large range of face shapes observable during communication:

neutral subject's neutral relaxed face.

u the "u" sound vowel.

o the "o" sound vowel.

i the "i" sound vowel.

csmile closed smile posture.

osmile mouth open smile posture.

clench open lips and clenched teeth posture.

open mouth wide open posture.

prot protrusion of the lip posture.

Once acquired, raw data are then processed to be ready for analysis: We manually draw the face feature lines on the 3D mesh (eyebrows, eyes, nose, lips and face contour) to extract only the subject face (eyes and lips are removed because they bring their own modelling problem). Then we adapt the raw mesh to a generic mesh model using feature-based metamorphosis [7] to reduce the number of points to 436 [3] shown in figure 1 . For the analysis, a subset of 120 subjects was extracted containing equal numbers (30 subjects each) of Japanese men, Japanese women, Caucasian men, and Caucasian women. All nine postures were used for each subject, giving a total of 1080 3D faces.

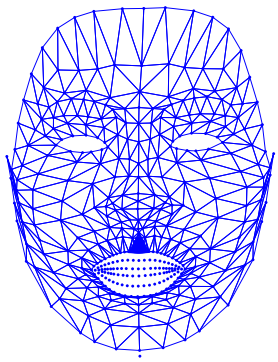

Figure 1: *Repartition of the 436 points of the adapted mesh.*

2. Principal Components Analysis of the database (PCA)

2.1. PCA principles

The first step of the analysis was to perform a PCA on the face data. So a matrix $F = [(f_i)]$ is formed, where the i^{th} column contains the 3D coordinates of the corresponding face. This matrix is zero centered by removing the mean then the faces' orientation is aligned to the mean face $F0 = align(F - mean)$. The PCA consists of finding an orthonormal base of vectors that best describes the covariance matrix $COV = F0 \times F0^T$. We obtain this base with a singular value decomposition :

$$\exists U, L \mid COV = U \times L \times U^T \quad (1)$$

where L is a diagonal matrix that contains the sorted eigenvalues of COV and U the orthonormal base of the eigenvectors. The i^{th} eigenvalue of COV tells us how much the i^{th} eigenvector of U contributes to recover the variance of the face cloud.

2.2. PCA results

This method gave good results and the components all gave realistic and coherent modifications of the face. Figure 2 plots the number of principal components needed to recover a certain percentage of the variance of the whole cloud of faces. One can

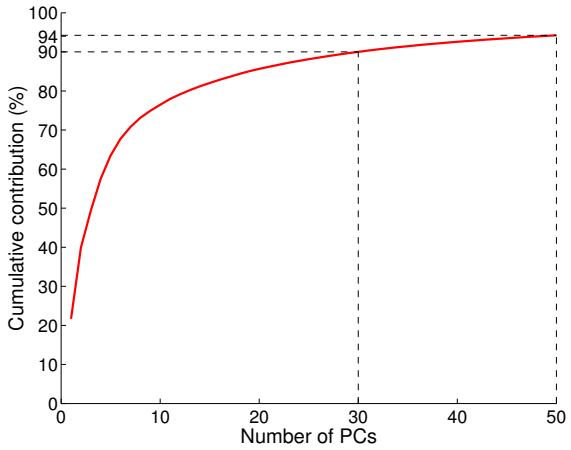

Figure 2: *Cumulative contribution of principal components (PCs) in **all** postures analysis. 30 PCs are necessary to recover 90% of the variance.*

notice that only 30 Principal Components (PCs) are needed to recover 90% of the variance of the face cloud and 94% with 50 PCs. In other words we can reasonably write a face f_m as a linear combination of the first N PCs as shown in equation 2 :

$$f_m = mean + \sum_{i=1}^{N} Z_{i,m} * U^i \quad (2)$$

U^i is the i^{th} column of U, N the number of components retained and Z is the projection of $F0$ on the new base $Z = U^T \times F0$. In this way, we can significantly reduce the dimensionality of the cloud from $436 * 3$ to N. In the following we used $N = 50$.

We performed several analyses including some or all of the postures and also including all postures. For example figure 3 shows some results of the *all* postures analysis: left of figure 3 shows the first PC at minus one standard deviation and right side shows $PC2$ at plus one standard deviation. We also built a real time visualization tool using sliders of the PCs called PCAview shown in figure 4 and it appeared that the major PCs have a meaningful influence on the face that could make it easy to separate face characteristics.

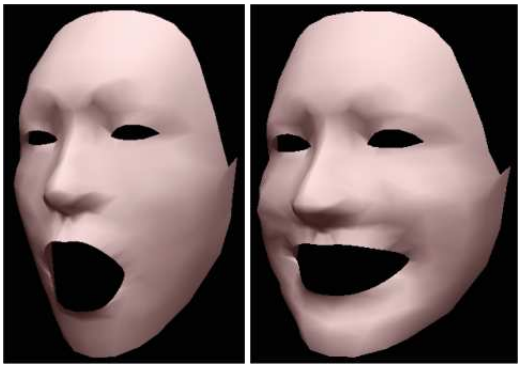

Figure 3: *PC1 at -1*standard deviation (left) and PC2 at 1*standard deviation (right) from the all postures analysis.*

Figure 4: *Slider based visualization tool PCAview. Screen capture example.*

It is clear that the PCs carry interpretable information such as gender or ethnicity according to figure 5. In a single PC, one can notice changes in apparent ethnicity and gender but for the smaller PCs (above $PC15$) we find it very difficult to interprete the influence on the face. Then comes the idea of performing the Linear Discriminant Analysis of different populations.

3. Linear Discriminant Analysis of the PCs coefficients

3.1. LDA principles

The Linear Discriminant Analysis is a recent technique for best separating two or more already known clouds of points. In our case, given two sets of faces, *Japanese* and *Caucasian*, the LDA gives the direction that maximizes the distance between the two clouds. LDA could have been performed on the 3D-vertices directly but the main concern was to get an automatic way of analyzing the PCs' influence on the face. The result of the LDA is a vector D_{pc} that contains the weights to use for each PC to

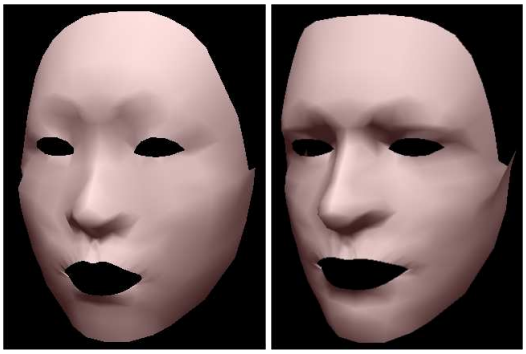

Figure 5: *neutral PC3 at minimum looks like a Japanese woman (left) and at maximum looks like a non-Japanese man (right)*

get the best separating direction D in the face space.

$$D = U \times D_{pc} \tag{3}$$

The LDA consists of the following calculation steps as presented in [4] : first we separate the PCs coefficients of the Z matrix into two matrices, J for Japanese faces and C for Caucasian faces. This LDA was performed with the neutral face of each subject only.

Then we define $\mu_J = \underset{i}{mean}(J^i)$, $\mu_C = \underset{i}{mean}(C^i)$ and p_J and p_C as the probability for a face to belong to J or C (we used 0.5). Then the mean of the entire set is $\mu = p_J * \mu_J + p_C * \mu_C$.

Let

$$\begin{aligned} cov_J &= (J - \mu_J) \times (J - \mu_J)^T \\ cov_C &= (C - \mu_C) \times (C - \mu_C)^T \end{aligned} \tag{4}$$

be the covariance matrix of each set.

From here, two types of analyses can be achieved : the class independent separation which gives one direction for the two groups and the class dependent separation which gives a direction for each group. We used the class independent transform to be able to gather the discriminated criteria on one direction only.

Therefore we define

$$\begin{aligned} S_w &= p_J * cov_J + p_C * cov_C \\ S_b &= (\mu_J - \mu) \times (\mu_J - \mu)^T + (\mu_C - \mu) \times (\mu_C - \mu)^T \end{aligned} \tag{5}$$

S_w being the within-class scatter and S_b the between class scatter.

The optimizing criterion in LDA is the ratio of between-class scatter to the within-class scatter :

$$criterion = S_w^{-1}.S_b \tag{6}$$

We have a 2-class problem so the criterion matrix can be reduced to one direction : the criterion first eigenvector. So we perform singular value decomposition on $criterion$ as in equation 1 to get the first eigenvector of PC coefficient D_{pc} of the new base. The final direction D is resynthesized with equation 3.

3.2. LDA performance

We noticed that the LDA performance depends mainly on two factors :

The standard deviation of the PCs coefficients. The PCs we obtained with the previous method follows $\|U\| = 1$. This implies that the weight (the eigenvalue information) of each components is held by the PCs coefficients. Another PCA technique would have given non normal eigenvectors but PCs coefficients with the same standard deviation. We have artificially normalized the PCs coefficients so that the LDA won't focus too much on the lower components so we get an LDA direction with a big spatial variance but a little sensitivity for population separation in that case.

The number of PCs used The more components we used the more information we can highlight. However we noticed an over-adaptating effect with a high number of PCs. This means that the LDA emphasizes information that is specific to the set we used. To evaluate this effect, we ran "jack-knife" tests by removing each face once from the set before analyzing and looking at the misclassification rate of this removed face with the number of PCs.

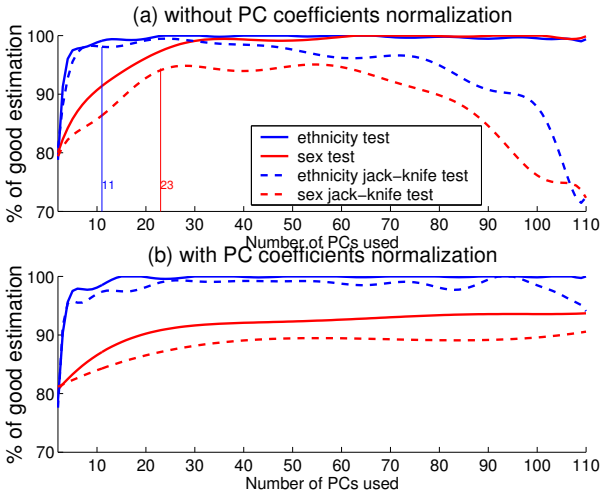

Figure 6: *(a) Without normalization the jack-knife test results emphasize the over-adaptating effect with a high number of PCs the best compromise are achieved with 11 components for ethnicity and 23 for gender.(b) With using the normalization of the PC coefficients, the performance are not as good as in (a), but we are sure that the major components appeared according to their importance. Then the LDA directions we get have a strong visual effect as seen in figure 9 and 10 but a less efficient separation.*

Figure 6 illustrates how to determine the best number of components to have good gender and ethnicity separation and strong spatial face variance which is equivalent to robustness. The best compromise is obtained with $N = 11$ for the ethnicity direction and $N = 23$ for gender.

We performed LDA on face gender and ethnicity and it appeared that with only a few PCs the predictions of the gender or ethnicity of a face reaches quickly more than 95% of success as shown in figure 6. Figure 6.(a) also shows that the ethnicity separation seems to be less ambiguous than the gender separation

and requires fewer components. Figure 7 shows the database faces distribution along the ethnicity and gender LDA direction obtained. As can be seen, there is clear separation between the groups and very few misclassifications.

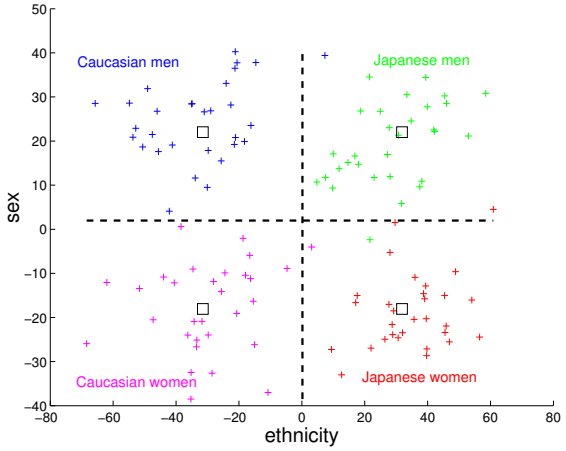

Figure 7: *Good population separation with LDA 11 PCs for ethnicity and 23 PCs for gender.*

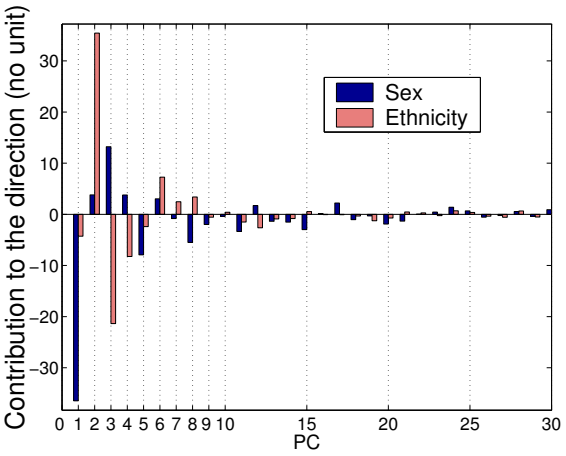

Figure 8: *Contribution of each PC from the neutral analysis to the LDA directions*

Looking at the direction, the gender and ethnicity information of the face is captured in the first few PCs as shown in figure 8 but an efficient and robust classification requires more. After PC15 their influence decreases very quickly. One can see that only the first eight PCs really contribute to the gender or ethnicity. Figure 9 and 10 show a face f rendered with PCAview as

$$f = mean + \alpha * D_{dir} \quad (7)$$

with $\alpha = f.D_{dir}$ being extremes values found in the database.

4. Linear face structure

The LDA results make it possible to build a 3D linear face model. We already have a vectorial space of PCs that fits the centered face space. Such a space is very easy to manipulate and addition, substraction and products between vectors (faces)

Figure 9: $dir = ethnicity$, $\alpha = -60$ *(left)*, $\alpha = 60$ *(right)*

Figure 10: $dir = sex$, $\alpha = -40$ *(left)*, $\alpha = 40$ *(right)*

easily make intuitive sense. We have in fact several vector spaces and the transformation from one to another are linear reversible operations if we consider the loss of data when reducing the number of PCs as negligible. The basic vector

Original space		PC space		Face model
Faces dimension 436×3	PCA $\Longrightarrow$	PC coef. dimension N	LDA $\Longrightarrow$	qualitative dimension 4

Figure 11: Vector space transforms

operator $+$ and $-$ can be used between faces, PC coefficients or LDA directions. The result of those operation remains in the vector space and correspond to addition or substraction of faces features. The Euclidean dot product $u.v = u^T \times v$ corresponds to a measure of orthogonality and the standard definition can be applied either to the faces or to the PC coefficients because the PC base is orthonormal.

The LDA directions, D_s for gender and D_e for ethnicity, have a norm of 1 so the dot product between them gives a good idea of how close they are. In the neutral posture analysis case the dot product between the ethnicity and gender directions $D_s.D_e = 0.04$ says that the directions are not absolutely orthogonal but have some part in common as seen in figure 8. This non perfect orthogonality is also illustrated in figure 9 and 10 where psychologically talking, each direction contains some information of the other. To reduce the dimensionality of the face space we can characterize a face through its gender,

ethnicity and posture.

Posture robustness.
It also appears that the ethnicity and gender estimation are independent of posture, shown in the table:

	neutral	u	i	o	clench
Ethnicity	98%	98%	98%	99%	98%
Sex	93%	94%	90%	91%	89%

	osmile	open	csmile	prot
Ethnicity	97%	98%	97%	97%
Sex	84%	89%	97%	95%

	mean
Ethnicity	98.2% $\sigma = 0.6$
Sex	90.7% $\sigma = 3.6$

Moreover, an LDA-like analysis, Multiple Discriminant analysis, showed that the nine postures information seems to be held at 95% by two directions named D_{post1} and D_{post2}. The same MDA technique applied to gender and ethnicity together gives the same result as LDA and confirms that we need two directions almost orthogonal in the mathematical sense (dot product < 0.1).

Now we have a simple approximative face model from the transformation described in figure 11

$$f = mean + \alpha_e * D_e + \alpha_s * D_s + \alpha_1 * D_{post1} + \alpha_2 * D_{post2} + \varepsilon_r \quad (8)$$

where $\alpha_e = f.D_e$, $\alpha_s = f.D_s$, $\alpha_1 = f.D_{post1}$, $\alpha_2 = f.D_{post2}$ and ε_r is the remaining information not accessible (orthogonal) through the gender, ethnicity or posture directions. ε_r contains the subject's special shape features. You can see a layer substraction process in figure 12.

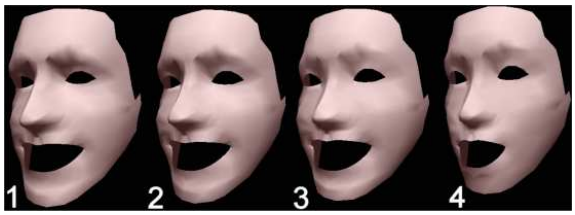

Figure 12: *1. original face, 2. face - eth, 3. face - gender - eth, 4. face - gender - eth - posture*

5. Comparison with human perception

In order to check if the LDA directions found in section 3.1 are more than a statistical interpretation of gender or ethnicity and have any psychological validity, we have run perception experiments. We showed 3D freely orientable faces shape with two rating scales, one that went from woman (1) to man (7) and the other from Japanese (1) to Caucasian (7).
We ran two different experiments :

1. In the first experiment we chose real faces from the database : 5 Japanese women, 5 non-Japanese women, 5 Japanese men, 5 non-Japanese men so that their distribution covered the gender and ethnicity directions. Then each face was shown four times : once the original face, once the face minus its ethnicity component, once the face minus its gender component and once minus its ethnicity and gender component (80 faces total). We run this test with 12 subjects balanced for gender and ethnicity.
2. The participants were shown a sequence of 60 faces synthesized from the first 30 Principal Components of the neutral posture analysis. The coefficients were randomly chosen between -2 and 2 times their standard deviation following a normal centered distribution.

First experiment results. (figure 13)
The first experiment showed that, consistent with the data shown in figure 6, the ethnicity is easy for participants to determine. It seems that the face shape carries more information on ethnicity than on gender. The performance for the two criteria is fairly good as seen in first part of figure 13, around 90% for ethnicity and 80% for gender. This validates the fact that the face shape is sufficient to make a reliable decision on a person's gender or ethnicity. The three other parts of figure 13 show that the information held in the LDA directions is perceptually salient. When we remove either the gender or ethnicity components on a face, participants can no longer make a reliable decision and their answer is about chance (50% correct). Moreover, it also confirms the orthogonality in the sense that removing one component does not affect the decision on the other.

Second experiment results.
The results for the second experiment show that decision making for randomly synthesized faces is no more difficult than for real faces, the artificial faces are considered as real faces. We call ideal observer, the decision that the computer makes using the database LDA directions that we will call D^{φ} for "physical". We can now measure how much the participants and the ideal observer agrees on the artificial faces. The participants'answers match the ideal observer decision at 63% for gender 71% and for ethnicity. This reflects a slightly different judging method. We then compute the LDA directions D^{ψ} out of the participants answers. This gives us a way to visually study the criteria used to make their decision, in other words what participants are looking at to decide. We saw that the psychological directions D^{ψ} are in all cases close to D^{φ}.
However the dot products are

$$D_e^{\varphi}.D_e^{\psi} = 0.9 \; and \; D_s^{\varphi}.D_s^{\psi} = 0.7 \quad (9)$$

This means that the criteria used to judge the ethnicity is almost identical for the ideal observer and the participants. But in the gender case the difference tends to be significant. This is to be linked with the fact that performance is lower, as shown in the first experiment. We also note that

$$D_e^{\varphi}.D_s^{\varphi} = 0.1 \; but \; D_s^{\psi}.D_e^{\psi} = 0.75 \quad (10)$$

Suggesting that we don't have orthogonality between the psychological gender and ethnicity directions. The participants differed from the ideal observer in judging gender with similar criteria as ethnicity. Once more, we can suppose that they felt that they had to make a decision they were not confident about doing so. But many other hypothesis are possible.

6. Implications and limitations

It is although necessary to be careful with the scope of those results. Equation 10 demonstrates a big difference between psychological and physical directions. The participants are using very close directions for deciding of gender and ethnicity

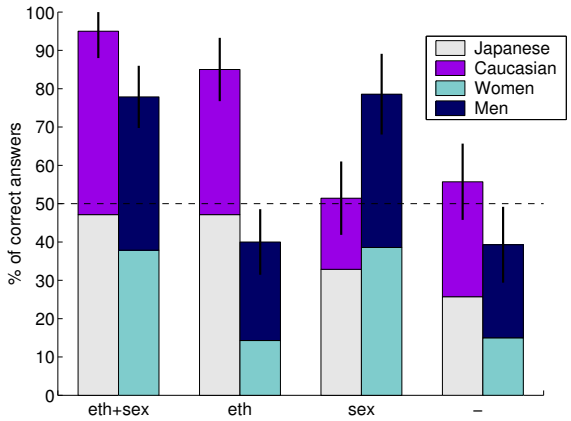

Figure 13: *The LDA directions contains the effective judging criteria. The abscise axe shows the information contained in the shown face.*

whereas we have seen that the physical directions are almost orthogonal. Combined with equation 9, we can conclude that D_s^{ψ} defines a direction with missing or added information. Several hypothesis have arisen :

1. The human brain is using a synthetic criteria to decide on gender that cannot be completely extracted with this statistical method.
2. The shape doesn't hold enough information to decide gender precisely for a face if it is mixed with ethnicity information. For example on both directions the Japanese women can be easily distinguished from Caucasian men but the distinction between Japanese men and Caucasian women is more subtle. The texture component is probably very decisive as seen in [5] but also complementary [6].
3. The mesh adaption process may add or lose some facial details at the reduced mesh resolution.
4. Figure 7 clearly shows that gender distribution behave differently from ethnicity distribution, we can think that generally speaking although the ethnicity is a less binary decision, the resemblance are more likely to happen between men and women than between Japanese and Caucasian.
5. According to figure 2, using 30 PC should retain most of the variance but what does the other 10% missing account for? As the low rank components induce small and difficult to interpret variations on the face, the dimensionality reduction has a vagueness cost.

Although it remains very difficult to measure, we think that hypothesis 3 must induce face distortions. Naturally, hypothesis 4 is obviously making the task more difficult because gender is more likely to be confused than ethnicity in the set we used.

7. Conclusion and discussion

The PCA technique gives an efficient low dimensional representation of the face shape with a psychological validity; random faces can be considered as valid. The LDA did an effective categorization of gender and ethnicity within this space as performance dropped when the directions components were removed. In future work, we will try to highlight a time adaption effect on perception and compare the different directions even within groups (differences in feminity in Japanese and Caucasian for example). We saw in this paper that with combining LDA and PCA, we could built a face shape linear model that can highlight independent ethnicity, gender and posture layers. Moreover those layers have a real psychological impact on perception. We can probably go on "removing" independent layers with this method such as age or even subdivide the ethnicity layer for example. We have an easy method to subtract or add meaningful face features and we can push further investigation on the last term ϵ_r of equation 8.

We can expect that beyond the posture, ethnicity and gender facial characteristics would appear some subject-specific features that could be useful for face recognition purpose. We have then access to a kind of "face print". But we are sure that this ϵ_r summarize much more relevant and non-specific information. One of the main goals of the face database is to be able to animate faces in a realistic way in terms of communicational information . Knowing that static ethnicity, gender and posture can be linearly added to an existing face makes avatar and cross-subject animation much simpler and reliable.

8. Acknowledgments

This research was supported by Telecommunications Advancement Organization (TAO) of Japan. Vatikiotis-Bateson was also supported in Canada by NSERC.

9. References

[1] Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion, and speech acoustics. J. Phonetics, 30, 555-568.

[2] Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. (1998). Eye movement of perceivers during audio-visual speech perception. Perception & Psychophysics, 60(6), 926-940.

[3] Takaaki Kuratate, Hani Yehia, Eric Vatikiotis-Bateson. "Kinematics-based synthesis of realistic talking faces". AVSP'98, Terrigal, Australia.

[4] S. Balakrishnama, A. Ganapathiraju. "Linear Discriminant Analysis - A brief tutorial". Institue for Signal and Information Processing. department of Electricaland Computer Engineering. Mississipi State University.

[5] Harold Hill, Vicki Bruce and Shigeru Akamatsu. "Perceiving the sex and race of faces: the role of shape and colour". Proc R. Soc. Lond. B (1995) 261, 367-373.

[6] Alice J. O'Toole, Tjomas Vetter, Nikolaus F Troje, Heinrich H Bülthoff. "Sex classification is better with three-dimensional head structure than with image intensity information". Perception 1997, volume 26, pages 75-84.

[7] Thaddeus Beier, Shawn Neely. "Feature-Based Image Metamorphosis". Siggraph'92.