Assessing facial beauty through proportion analysis by image processing and supervised learning

Hatice Gunes * and Massimo Piccardi

Computer Vision Research Group, Faculty of Information Technology, University of Technology, Sydney (UTS)

PO Box 123, Broadway, NSW, 2007, Australia

Abstract

Perception of universal facial beauty has long been debated amongst psychologists and anthropologists. In this paper, we perform experiments to evaluate the extent of universal beauty by surveying a number of diverse human referees to grade a collection of female facial images. Results obtained show that there exists a strong central tendency in the human grades, thus exhibiting agreement on beauty assessment. We then trained an automated classifier using the average human grades as the ground truth and used it to classify an independent test set of facial images. The high accuracy achieved proves that this classifier can be used as a general, automated tool for objective classification of female facial beauty. Potential applications exist in the entertainment industry, cosmetic industry, virtual media, and plastic surgery.

Key words: facial beauty classification, observer agreement, proportion analysis, golden proportion, facial thirds, facial features, supervised learning, performance evaluation

1 Introduction

It has long been believed that the concept of facial beauty is variable and subjective to race, culture or era. However, psychological and medical sciences state that there is a timeless, aesthetic ideal facial beauty based on

* Corresponding Author.
Email address: haticeg@it.uts.edu.au (Hatice Gunes).

facial proportions. Recent scientific studies also reveal that the concepts of a "beautiful face" are not learned but seem to be "hard-wired" into our mind from birth (Larrabee, 1997), (Yellin, 1997). Moreover, cross-cultural investigations on facial beauty show that different groups have similar perceptions of facial beauty, and a universal concept of beauty may be defined throughout different races, cultures and eras (BBC Science, 2002), (Bell, 1997), (Cunningham, Roberts, Barbee, Druen, et al., 1995), (Daibo, 1999), (Epker and Koury, 1992), (Farkas, 1985), (Farkas, 1987), (Farkas, 1994), (Jefferson, 1993), (Landau, 1989), (Langlois and Roggman, 1990), (Larrabee, 1997), (Marquardt Beauty Analysis, 2002), (Mealey, Bridgstock and Townsend, 1999), (Michiels and Sather, 1994), (Parris and Robinson, 1999), (Perrett, May and Yoshikawa, 1994), (Ricketts, 1982), (Yellin, 1997).

Attempts at measuring beauty quantitatively have been made by investigators in psychology, arts and image analysis, and more recently in oral and maxillofacial surgery (BBC Science, 2002), (Bell, 1997), (Cunningham, Roberts, Barbee, Druen, et al., 1995), (Daibo, 1999), (Epker and Koury, 1992), (Farkas, 1985), (Farkas, 1987), (Farkas, 1994), (Jefferson, 1993), (Landau, 1989), (Langlois and Roggman, 1990), (Larrabee, 1997), (Marquardt Beauty Analysis, 2002), (Mealey, Bridgstock and Townsend, 1999), (Michiels and Sather, 1994), (Parris and Robinson, 1999), (Perrett, May and Yoshikawa, 1994), (Ricketts, 1982). The most famous of these are based on the Golden Proportions (derived from the Golden Ratio, or phi: 1.61803) (BBC Science, 2002), (Borissavlievitch, 1958), (Huntley, 1970) and the Facial Thirds (Farkas, 1985), (Farkas, 1987), (Farkas, 1994). However, measuring facial beauty remains a challenging task. The existing approaches either lack general confirmation from a significant pool of human referees, or require several cumbersome manual measurements, or both. Instead, it is important that the approach be based on experiments with sufficient human referees and automated image analysis tools. Such an automated and objective beauty classifier could be extremely useful in several applications such as plastic surgery (for predictive evaluation of facial beauty before surgical procedures), the cosmetic and entertainment industries, and virtual media. In the following, we briefly review the main methods from the recent literature and introduce our approach thereafter. The group from (Marquardt Beauty Analysis, 2002) has developed a method that measures facial beauty by using the Golden Ratio. Although they proved the usefulness of proportions in facial features, their method does not carry documented evidence of its validity; moreover, it requires several manual measurements. Aarabi and Hughes in (Aarabi and Hughes, 2001) proposed an automated procedure based on the data from 12 human referees and k-nearest neighbor classification. Their grading system, however, only provides coarse classification into 4 beauty classes, thus limiting usefulness for many real applications. Moreover, the set of human referees was neither sufficiently large nor diversified. Instead, an automated procedure was proposed in (Gunes and Karsligil, 2002) based on automated image analysis and a rule-based classi-

Fig. 1. System Framework.

fication system. The classification rules were elicited by hand and they still required validation. In order to improve both validation and automation of the procedures, in this paper, we propose an approach providing the following advantages:

- the ground truth is based on a survey of tens of diverse human referees giving beauty grades to a collection of 215 female facial images of different age and ethnicity for a total of more than 10,000 grades. Such values are significantly larger than those of the previous approaches. The statistical results of this survey form our ground truth. The survey gives evidence that the grade histograms (the histograms of grades, or marks, for a same image from many human referees) are shown to be unimodal and compact in most cases, thus supporting the statement of "universal" facial beauty perception;
- the facial features of a face image are extracted automatically by using accurate image analysis operators;
- the extracted facial image features and the ground-truth classification are used together to train a tree-based classifier which provides automated grading of the facial beauty. The classifier is tested with an independent set of facial images, reporting high classification accuracy; this validates our system to be an objective, automated tool for classification of female facial beauty. Extension to male cases would be straightforward.

The system framework illustrating these steps is shown in Figure 1. The rest of the paper is organized as follows: Section 2 describes the facial proportions used for measuring facial beauty. Section 3 describes the face detection process and the feature extraction operators. Section 4 describes the human classification and automatic classification process. Section 5 gives an evaluation of the performance of the automated classifier and, finally, Section 6 presents the conclusions and future work.

2 Facial proportions

In this section we present the facial proportions used in this study specifically for the vertical and horizontal sets (Epker and Koury, 1992), (Farkas, 1985), (Farkas, 1987), (Farkas, 1994), (Marquardt Beauty Analysis, 2002), (Parris and Robinson, 1999).

Fig. 2. The Golden Ratio.

2.1 Vertical proportions

The vertical proportions we considered are based on two main aesthetic theories: the Golden Proportion and Facial Thirds methods (Epker and Koury, 1992), (Farkas, 1985), (Farkas, 1987), (Farkas, 1994), (Marquardt Beauty Analysis, 2002), (Parris and Robinson, 1999).

Golden Proportions. The Golden Ratio or Proportion is approximately the ratio of 1 to 0.618 or the ratio of 1.618 to 1 (Borissavlievitch, 1958), (Huntley, 1970) as shown in Figure 2. According to the Golden Proportions method for female facial beauty, in the case of a perfect, vertically aligned face, all the proportions stated in Table 2 must fit the Golden Ratio (Parris and Robinson, 1999). The description of the distances that we used for the Golden Proportions ratio calculations are given in Table 1 and also illustrated in Figure 3(a). Table 2 describes the specific ratios used in the Golden Proportions method and by our system. It is easy to see that there are other ratios that might be derived as a consequence of the Golden Proportion Rule and other ratios (such as for instance (3:4)) respect the Golden Proportions. However, they were not used in the original work from (Parris and Robinson, 1999) as they are redundant and as such not used in this work either.

Facial Thirds. The Facial Thirds method states that a well-proportioned face may be divided into roughly equal thirds by drawing horizontal lines through the forehead hairline, the eyebrows, the base of the nose, and the edge of the chin (see Figure 3(b)). Moreover, the distance between lips and chin should be double the distance between the base of the nose and lips (Farkas, 1985), (Farkas, 1987), (Farkas, 1994), (Marquardt Beauty Analysis, 2002), (Parris and Robinson, 1999). Table 3 summarizes the Facial Thirds ratios.

(a)

(b)

(c)

Fig. 3. Template images for (a) Golden Proportions; (b) Facial Thirds; (c) horizontal proportions.

Golden Proportions and Facial Thirds methods are similar to each other. However, the former specifies a larger number of proportions than the latter. Moreover, they use different measuring points and as such cannot be directly compared. However, if we align them to the same measuring points, they seem in good accordance. For instance, taking into account Figure 4 let us consider the following two ratios:

Table 1
Description of the distances used when defining the vertical facial ratios.

1	vertical distance between top of the face and tip of the chin
2	vertical distance between pupils and tip of the chin
3	vertical distance between top of the face and nostrils
4	vertical distance between top of the face and pupils
5	vertical distance between nostrils and tip of the chin
6	vertical distance between pupils and central lip line
7	vertical distance between lips and tip of the chin
8	vertical distance between pupils and nostrils
9	vertical distance between nostrils and central lip line
10	vertical distance between top of the face and eyebrows
11	vertical distance between eyebrows and tip of the nose
12	vertical distance between tip of the nose and tip of the chin
13	vertical distance between tip of the nose and lips
14	vertical distance between lips and tip of the chin

Table 2
Ratios used in the Golden Proportions method (Parris and Robinson, 1999).

2:4	vertical distance between pupils and tip of the chin to vertical distance between top of the face and pupils
3:5	vertical distance between top of the face and nose to vertical distance between nostrils and tip of the chin
6:7	vertical distance between pupils and central lip line to vertical distance between lips and tip of the chin
5:8	vertical distance between nostrils and tip of the chin to vertical distance between pupils and nostrils
8:9	vertical distance between pupils and nostrils to vertical distance between nostrils and central lip line
7:9	vertical distance between lips and tip of the chin to vertical distance between nostrils and central lip line
mean_ratio	mean of the six aforementioned ratios

Table 3
Ratios used in the Facial Thirds method.

10:1	Ratio of vertical distance between top of the face and eyebrows to face length
11:1	Ratio of vertical distance between eyebrows and tip of the nose to face length
12:1	Ratio of vertical distance between tip of the nose and tip of the chin to face length
13:14	Ratio of vertical distance between tip of nose and lips to vertical distance between lips and tip of the chin

Fig. 4. Comparison between measurement points of Golden Proportions and Facial Thirds.

Table 4
Description of the distances used when defining horizontal facial ratios.

1	vertical distance between top of the face and tip of the chin
15	horizontal distance between center of the pupils (inter-eye distance)
16	horizontal distance between the widest parts of the face

1:3 Length of the face to the distance between top of the face and the nostrils;

1:X Length of the face to the distance between top of the face and tip of the nose;

According to the Facial Thirds rules, *(1:3)* should be equal to 1.5 in an ideal case and according to the Golden Proportion rules *(1:X)* should be 1.618. Figure 4 shows that *segment 3* is slightly shorter than X for any face and, as a consequence, *(1:3)* is always slightly smaller than *(1:X)*. This justifies the small difference in the numerical values from the two sets of rules showing their substantial agreement.

2.2 Horizontal proportions

For horizontal proportions, we propose to use the ratio of "the inter-eye distance to face width" as the eyes are the most notable horizontal landmark in a face and the ratio of "facial width to face lenght" to capture the proportionateness of the face elongation. For detailed description of such horizontal ratios, please see Figure 3(c), and Tables 4 and 5.

According to these methods, a face is more attractive as it approaches the

Table 5
Ratios used in the Horizontal Proportion calculation.

15:16	Ratio of inter-eye distance to face width
16:1	Ratio of face width to face length

proportions shown in Figure 3(a), 3(b) and 3(c), respectively. Facial plastic surgeons have already been using these ratios as a guide for their work (Bell, 1997), (Cunningham, Roberts, Barbee, Druen, et al., 1995), (Daibo, 1999), (Epker and Koury, 1992), (Farkas, 1985), (Farkas, 1987), (Farkas, 1994), (Jefferson, 1993), (Landau, 1989), (Langlois and Roggman, 1990), (Larrabee, 1997), (Marquardt Beauty Analysis, 2002), (Mealey, Bridgstock and Townsend, 1999), (Michiels and Sather, 1994), (Parris and Robinson, 1999), (Perrett, May and Yoshikawa, 1994), (Ricketts, 1982), (Yellin, 1997). Computing all the aforementioned proportions requires accurate feature extraction, that we perform in six phases: face localization by skin region detection, eye localization, pupil localization, eyebrow localization, base of the nose localization, lip localization and, finally, chin localization.

3 Feature extraction and analysis

Accurate feature extraction is of fundamental importance for reliable measurement of facial beauty. A vast literature covers techniques for facial feature extraction (Hjelmas, 2002), (Yang, Kriegman and Ahuja, 2002), (Rein-Lien, Abdel-Mottaleb and Jain, 2002), (Graf, Chen, Petajan and Cosatto, 1995), (Yow and Cipolla, 1995), (Yow and Cipolla, 1996), (Yow and Cipolla, 1997). In this work, we choose to use the well-known methods proposed in feature-based face detection approaches since such methods have proven reliable and computationally efficient (Hjelmas, 2002), (Yang, Kriegman and Ahuja, 2002), with main references to (Graf, Chen, Petajan and Cosatto, 1995), (Yow and Cipolla, 1995), (Yow and Cipolla, 1996), (Yow and Cipolla, 1997).

The first step consists of locating the facial region in order to remove irrelevant picture information. We assume that the largest part of each image is the portrait view of a face and that each face is located in the center of the image. In our test set, the center of the image will identify the face-skin color of the observed person. Hence, we calculate the average pixel value from the center of the gray facial image, within a 7 * 7 pixel region, assuming that it corresponds to the facial region. Starting from this region, we then apply a region growing step to improve detection of the contour of the facial region. The resulting image from this step is shown in Figure 5(a).

After identifying the facial region, the eyes are usually detected first. After band pass filtering, morphological operations are applied to enhance regions

Fig. 5. (a) Facial region after region growing (enhanced and binarized image); (b) Horizontal position of eyes located as histogram local minimum in the upper histogram part; (c) Resulting image from horizontal edge detection; (d) Location of the connected components identified as eyebrows; (e) Location of the lip line; (f) Horizontal locations of eyebrows, pupils, tip of the nose, lip line and tip of the chin.

Fig. 6. Input image and the corresponding histogram.

with high intensity (e.g., eyes). The histogram of the enhanced image typically exhibits a prominent peak. Based on the peak value and its width, adaptive threshold values are selected in order to generate binarized images. For eye detection, the vertical histogram of the skin-region is computed as shown in Figure 6. The rows containing the eyes are located in correspondence with a histogram local minimum in the upper histogram part (see Figure 5(b)). Furthermore, the horizontal histogram for these rows is also computed, and eyes are then located as the two local minima. Finally, connected component labeling is applied to the binarized image in order to identify the areas of candidate eye regions.

Furthermore, we exploit the knowledge about facial anatomy and use the facial model and partial face groups introduced by Yow and Cipolla when detecting the eyes and eyebrows (Yow and Cipolla, 1996). They model the face as a plane with 6 oriented facial features (namely the eyebrows, the eyes, nose and mouth) and use a second derivative Gaussian filter, elongated at an aspect ratio of 3 to 1, to detect interest points. Such points, detected at the local maxima in the filter response, indicate the possible locations of facial features.

The second stage examines the edges around these interest points using edge detection (i.e. Canny edge detector (Canny, 1986) in our case) and groups them into regions (see Figure 5(c)). The labeled features are further grouped based on the model knowledge of where they usually occur with respect to each other (Hjelmas, 2002).

After detecting the pair of possible eyes satisfying the geometrical constraints imposed by the human face, it is easier to localize the eyebrows. Eyebrows are expected to be located in the upper part of the face and are the first non-skin components on the facial region below the forehead as shown in Figure 5(d). Nose localization takes place after having located the eyes since its horizontal position is below the eyes and above the lips and its vertical position is in-between those of the eyes (Hjelmas, 2002). The location of the nose is later refined after having detected the lips. Lips anatomically are located between the eyes and below the nose; lips can be easily discriminated from skin based on their different intensity levels (Hjelmas, 2002), (Yang, Kriegman and Ahuja, 2002). Searching for the lips is relatively straightforward thanks to the well-

Table 6
Age and gender information of human referees.

	Age 18-25	Age above 25
Female	10	3
Male	17	18

Table 7
Background information of human referees by continental area.

Ethnicity	Number
Africa	1
South East Asia	8
India	2
Australia	2
Eastern Europe	10
Western Europe	15
Europe (unspecified)	4
North America (USA)	2
Latin America	4

Fig. 7. Marking scheme used in the survey (the image used is one of the authors'; display of a real face from the database is omitted for privacy reasons).

confined search space limited by nose and pupils (see Figure 5(e)). The chin detection process takes place after lip detection, as the chin anatomically is located between the lips and the neck. Hence, the search space is arranged according to the lip line and the horizontal lower limit of the facial region (see Figure 5(f)).

4 Human classification of facial beauty

In this work, we conducted a survey on a collection of female facial images with 48 human referees of different age and ethnical background. We organized a set of 215 female face images varying in ethnicity and age and asked the human referees to score them according to a scale of n = 10 grades, or marks, labeled into 5 categories (1: least attractive - minimum; 10: most attractive - maximum). For details on the pool of human referees, please see Tables 6 and 7. An example of the grading system is shown in Figure 7.

It is arguable that there may be differences in interpretation of the marking scheme used to mark the beauty level. However, there are some inherent difficulties in marking schemes in general (Blumhof and Stallibrass, 1994). These difficulties include:

- tendency to mark the more immediate concepts;
- tendency to mark towards the middle to avoid obvious offence (even if the survey is double-blind);
- exposing the subjectivity of marking schemes by trying to decide on, and weight, criteria. For instance, a mark of seven might represent a high mark for one referee, whereas the same mark for another referee might represent a concept of just above average.

For the purpose of this experiment, all the referees were informed of the basic grading mechanism. In our survey, we used a scale of 10 grades by labeling grades with descriptive adjectives such as "beautiful", "attractive", and "unattractive" to try to minimize the ambiguity in the marking scheme.

4.1 Analysis of Grade Distributions

In the following, we assume that each mark, x, belongs to a step scale with a unit interval, or grade class, [x - 0.5, x + 0.5). As such, marks for each I_i image from all the referees constitute a grade histogram for that image. We tested the unimodality of the mark distributions for each image based on the test defined in (Bottomley, 2004). By applying the Bottomley test to our case, the distribution proves unimodal if *the maximum distance between the mode and the mean* is bounded as follows:

$$|M_i - \mu_i| \ < \ \sqrt(3)\sigma_i \tag{1}$$

M_i: mode of the marks for image I_i
μ_i: average of the marks for image I_i
σ_i: standard deviation of the marks for image I_i

We computed the aforementioned measure for all 215 images and found that, for all of them, the marks fit a unimodal distribution.

Having passed the unimodality test, we then approximated the marks for each I_i image with a probability density function $p_i(x)$ of μ_i average and σ_i standard deviation. Figure 8 shows an example of grade histogram (Image no. 82 from our data set). Figure 8 shows that a well-formed mode centred around 5 exists. However, it also shows that a few referees marked as different as 1 and 9. Therefore, when we put forward the statement that beauty is universal,

Fig. 8. An example of grade histogram for image 82 from the data set.

Fig. 9. The histogram of the standard deviations for 215 images.

we mean that as a statistical statement. In other words, the judgement is not identical. Yet, a strong *central tendency* exists.

The objective of the next experiment was that of measuring the level of human agreement in grading the image set. In the case of complete agreement amongst the referees, all grades for each image would be exactly the same. Instead, different judgements would give place to dispersed grades. Moreover, different individual interpretations of the marking scheme might introduce a certain amount of dispersion even at a parity of judgement. In order to study the grade dispersion over the whole image set, we decided to compute the standard deviation of grades for each image, σ_i, and study their probability distribution, $\sum(\sigma)$, on the image set. Figure 9 shows the histogram of the standard deviations computed with step 0.1 on the set of 215 images. In order to explore the unimodality of the histogram we applied the test described in Equation 1. According to the test's result the histogram is unimodal, has a pronounced mode around $\sigma = 1.7$, is maximally symmetric and of approximately Gaussian shape, with a limited skew to the right. Therefore, we quantified the skewness to see if it was significant (Evans and Kuenning, 2002). The suggested measure is the standardised third sample moment around the mean:

$$\alpha_3(x_i) = \frac{\sum\limits_{i=1}^{N} (x_i - \mu)^3}{N\ \sigma^3} \tag{2}$$

where x_i, i = 1, ..., N are the samples and μ and σ the sample mean and standard deviation, respectively. If the distribution is symmetrical or non skewed, this quantity will be close to zero as positive and negative deviations will cancel each other. If the distribution is right-skewed, the cubed positive deviations outweigh the cubed negative ones yielding an overall positive number. Conversely, if the distribution is left-skewed, the cubed positive deviations outweigh the cubed negative deviations.

(Evans and Kuenning, 2002) state that a distribution with skewness less then 0.5 can be well approximated by a Gaussian distribution. In our case, the skewness of the standard deviations, $\alpha_3(\sigma_i)$, is equal to -0.0459. Since this value is close to zero we can conclude that the distribution of the standard deviations is not significantly skewed. In this case, the distribution can be appropriately described by its mean, μ_σ, and standard deviation, σ_σ. The smaller μ_σ, the higher the agreement on the perception of beauty over a set of different referees and a collection of different images.

Fig. 10. A Gaussian curve of standard deviation 1.628 superimposed to a grade class.

Table 8
Probability distribution of the standard deviation of grades, ($\sum(\sigma)$, over the set of images.

Average (μ_σ): 1.628

Standard deviation(σ_σ): 0.234

$$\mu_\sigma = \frac{1}{N_{img}} \sum_{i=1}^{N_{img}} \sigma_i, \quad \sigma_\sigma = \sqrt{(N_{img} \sum_{i=1}^{N_{img}} {\sigma_i}^2 - (\sum_{i=1}^{N_{img}} \sigma_i)^2)/{N_{img}}^2}$$

$$\sigma_i = \sqrt{(N_{ref} \sum_{j=1}^{N_{ref}} {x_{ij}}^2 - (\sum_{j=1}^{N_{ref}} x_{ij})^2)/{N_{ref}}^2}, \quad i = 1, ..., N_{img}$$

$x_{ij} \in \{1, 2, ..., N\}, j = 1, ..., N_{ref}$: grades of the human referees for the I_i image

Number of images (N_{img}): 215

Number of referees (N_{ref}): 48

Number of classes (N): 10

Table 8 shows that in this experiment the standard deviation in the human referees' grades is an average of $(\mu_\sigma) = 1.628$, less than twice the grade class size (which is unitary), thus proving that in general the agreement between the different human referees is high, even if not complete, and can be considered as the ground truth for the second part of our paper where we attempt at automatising the beauty assessment. To represent this graphically, Figure 10 shows a Gaussian distribution of standard deviation 1.628 superimposed to a grade class. Although this result is rather stable over the image set (σ_σ is low - see Table 8), we also found that the value of σ_i tends to be lower for images from most and least attractive faces, while for images with intermediate score the histogram tends to be less compact i.e. showing less agreement between the human referees.

4.2 Converse Probability Analysis

In this sub-section we aim to estimate the converse probability i.e. which variables, or factors, could be possible explanations for the distribution of the human votes. The variables we chose to analyze are: age, ethnicity and gender. Age and gender are well-known confounder variables in social and clinical trials and thus need to be explored. Given the potential influence of cultural issues in beauty judgment, we also decided to consider an ethnicity variable. Accordingly, we divide the grades of the referees into the following

Table 9
Mean, variance and standard deviation for groups of referees based on age, ethnicity and gender over the set of images.

variable		**mean of the population**	**variance of the population**	**stdev of the population**
age	18-24	4.48592	2.12803	1.45877
	25 and above	4.2846	2.46905	1.57132
ethnicity	European	4.44861	2.37074	1.53972
	Non-European	4.26935	2.26489	1.50495
gender	Female	4.79817	2.14504	1.46459
	Male	4.22179	2.3925	1.54677

group pairs: *age* (18-24/25 and above), *ethnicity* (European/ non-European) and *gender* (female/male). If results tend to remain stable for different values of these factors, we will be able to conclude that they do not significantly affect the score distribution. We demonstrate this hypothesis in two steps. Firstly, in Table 9 we provide descriptive statistics of the grade distribution for the three factors. A necessary condition to the claim for "universality" of beauty perception is that any group would show a similar mean and standard deviation. Table 9 shows that means and standard deviations are not identical, but actually very similar between the different groups. Therefore, none of these factors seem first-order explanations for the distribution of the human votes. Secondly, we decided to run hypothesis tests to probe these findings. The first test is a *paired t-test* assessing the statistical similarity or difference between the means of two groups of paired data. To obtain the paired data, for each I_i image, we divided the grades from the referees into two groups, 1 and 2, and computed their averages, x_{1i} and x_{2i}. We computed the t-test over such paired data to test their means, μ_1 and μ_2, over the whole set of images. The null hypothesis, H_0, that we formulate here is that the two means differ by a small Δ amount, chosen as the difference between the sample means of the two groups over the whole image set reported in Table 9. The alternate hypothesis, H_a, is that their mean difference is different from Δ:

$$H_0 : \mu_1 - \mu_2 = \Delta$$
$$H_a : \mu_1 - \mu_2 <> \Delta$$

The null hypothesis is rejected in favour of the alternate hypothesis if the t statistic:

Table 10
Paired t-test results on the means of referees' groups based on age, ethnicity and gender variables.

variable	Δ	df	t statistic	$t_{critical\,two\,tailed}$	H_a	H_0
Age	0.2	214	0.04547	1.97111	REJECTED	ACCEPTED
Ethnicity	0.18	214	-0.02408	1.97111	REJECTED	ACCEPTED
Gender	0.57	214	0.18011	1.97111	REJECTED	ACCEPTED

$$t = (\mu_1 - \mu_2)\sqrt{\frac{N_{img}(N_{img}-1)}{\sum_{i=1}^{N_{img}}((x_{1i}-\mu_1)-(x_{2i}-\mu_2))^2}} \qquad (3)$$

is either greater than a value called $t_{critical\,two\,tailed}$ computed on the t distribution, or smaller than $-t_{critical\,two\,tailed}$. Hypothesis tests also require to choose a *significance*, or *alpha*, level. Here we choose the typical alpha level of 0.05. Test results are shown in Table 10. In all cases, the null hypothesis is not rejected. Therefore, the small differences between the sample means of any two groups are statistically confirmed. It can be observed that:

- The difference between the means of any two groups is always significantly smaller than 1 which is equivalent to say that they differ by less than one mark.
- The highest difference between the means of two groups, 0.57 i.e. 5.7% of the grade range, is for the gender variable. This gives evidence to the slightly higher scoring provided by female referees with respect to that of male referees.
- The lowest difference between the means of two groups, 0.18 i.e. 1.8% of the grade range, is for the ethnicity variable.

For each pair of groups, we also applied an *F-test* to assess the statistical similarity or difference between the standard deviations of the two groups. We formulated the null and alternate hypotheses as follows:

$H_0 : \sigma_1 = \sigma_2$
$H_a : \sigma_1 > \sigma_2$

where σ_1 and σ_2 are the sample statndard deviations of the two groups. We ordered the two groups so that $\sigma_1 > \sigma_2$ so as to use a one-tail version of the F-test which is more convenient given the asymmetry of the F distribution. The alternate hypothesis is accepted if the F statistic:

Table 11
F-test results for the variances of the referees' groups based on age, ethnicity and gender factors.

variable	df	F statistic	$F_{critical\ one\ tailed}$	H_a	H_0
Age	214	1.16025	1.25279	REJECTED	ACCEPTED
Ethnicity	214	1.04673	1.25279	REJECTED	ACCEPTED
Gender	214	1.11536	1.25279	REJECTED	ACCEPTED

Table 12
The probability of a particular group given a particular set of grades with the age criterion (younger vs. older).

Score	p(x)	p(x | younger)	p(x | older)
(1-3)	0.3478	0.3169	0.3717
(4-6)	0.4569	0.4773	0.4410
(7-10)	0.1436	0.1350	0.1502

$$F = {\sigma_1}^2/{\sigma_2}^2 \tag{4}$$

is greater than a value called $F_{critical\ one\ tailed}$ computed on the F distribution. Again, the alpha level was set to 0.05. Test results are shown in Table 11. In all cases, the alternate hypothesis is rejected and the null hypothesis accepted confirming that variances are similar between any two groups.

Further to testing similarities of means and standard deviations, we analysed the probability of picking a particular group given a particular set of grades. In other words, we explored which variables, if any, could be possible explanations for the distribution of the human grades. We divided the grades of the referees into the following three sets: low (1-3), average (4-6), and high (7-10). We provide results in Tables 12, 13 and 14. Results for variables age and ethnicity show that conditional probabilities $p\,(group \mid set\ of\ grades)$ are similar for both groups and thus similar to the marginal probability $p\ (group)$. Results for variable gender show that the conditional probability for low grades, $p\,(group \mid (1-3))$ is lower for the female group than the male one. Conversely, the conditional probability for high grades, $p(group \mid (7-10))$ is higher for the female group. This confirms that the female referees tended to provide slightly higher grades than the male referees, also explaining the small differences in their average marks reported in Table 9.

Table 13
The probability of a particular group given a particular set of grades with the ethnicity criterion (European (E) vs. non-European (NE)).

Score	p(x)	p (x | NE)	p(x | E)
(1-3)	0.3478	0.3607	0.3385
(4-6)	0.4569	0.4805	0.4400
(7-10)	0.1514	0.1353	0.1628

Table 14
The probability of a particular group given a particular set of grades with the gender criterion (female vs. male).

Score	p(x)	p(x | female)	p(x | male)
(1-3)	0.3478	0.2640	0.3789
(4-6)	0.4569	0.4705	0.4518
(7-10)	0.1514	0.1843	0.1391

Fig. 11. Distribution of the individual features over the training set of 165 images.

5 Automatic Classification of Facial Beauty

As the next step, we aim to generate an automated classifier capable of reproducing the average human judgement of facial beauty by using a feature set based only on features automatically extracted from the images. We use the average grade, μ_i, given to the facial images by the human referees as our ground truth. In the following, we describe our approach for generating the automated classifier.

First, features extracted from each face were stored as tuples of attributes characterizing the face. The feature set contains several facial ratios as shown in Table 15. In addition, we show the histogram of each individual feature and that of the ground truth for a set of 165 images in Figure 11. Individual features do not show obvious clusters, thus supporting the idea that we can use an arbitrary number of grades as our classes. For classification purposes, several different methods including SVMs, neural networks, or other probabilistic models could be employed. We have chosen to use a supervised symbolic classifier, C4.5, based on the notion of entropy since its output - a decision tree - can be easily understood and interpreted by humans (Blumhof and Stallibrass, 1994), (Michalski, Carbonell and Mitchell, 1984), (Quinlan, 1993). Supervised classification requires the user to pre-classify a set of samples and train the classifier based on those samples. The training set must be large and varied enough to provide sufficient classifier training. Once our model is built, we

Table 15
The feature set for automated beauty classification.

From the Golden Proportion theory:	**From the Facial Thirds theory:**
2:4 vertical distance between pupils and tip of the chin to vertical distance between top of the face and pupils	**10:1** Ratio of vertical distance between top of the face and eyebrows to face length
3:5 vertical distance between top of the face and nose to vertical distance between nostrils and tip of the chin	**11:1** Ratio of vertical distance between eyebrows and tip of the nose to face length
6:7 vertical distance between pupils and central lip line to vertical distance between lips and the tip of the chin	**12:1** Ratio of vertical distance between tip of the nose and tip of the chin to face length
5:8 vertical distance between nostrils and tip of the chin to vertical distance between pupils and nostrils	**13:14** Ratio of vertical distance between tip of the nose and lips to vertical distance between lips and tip of the chin
8:9 vertical distance between pupils and the nostrils to vertical distance between nostrils and central lip line	**Horizontal proportions:**
7:9 vertical distance between lips and tip of the chin to vertical distance between nostrils and central lip line	**16:1** Ratio of face width to face length
mean_ratio (of the above)	**15:16** Ratio of inter-eye distance to face width

Fig. 12. The generated decision tree (sections).

need to estimate its accuracy. Estimating the accuracy of a classifier using training data could be optimistic and, thus, misleading. An analysis of accuracy tests for classification algorithms has been proposed in the much-cited (Dietterich, 1998). In the following sub-sections we discuss how we validate our approach with (a) the holdout method and (b) k-fold cross-validation.

5.1 Validation with the holdout method

For this validation method, we partitioned the 215 images into two independent datasets, a training dataset of 165 images and a test dataset of 50 images. Initially, we trained the classifier with 165 randomly selected images from our database. C4.5 (with parameters: confidence factor= 0.25, minNumObj= 1, numFolds=3) generated a decision tree of size 129 with 65 leaves. We then used a testing set of 50 images to test it. Analyzing the decision tree's structure can provide us with useful insights on the classification criteria. However, as the tree size is large, we report only some sections of it in Figure 12.

Several tests in the decision tree appear in good accordance with the aesthetics theories on proportions. For instance, in Section 1 in Figure 12, the attribute tested just prior to the leaf level (Ratio_(2:4), bottom two lines) is tested in agreement with the Golden Proportions rule: given its theoretically ideal value of approximately 1.618, face images with a value less or equal to 1.667 are given a grade of 7, while those with values above are given the much lower grade of 2. Section 2 describes the full path leading to the leaf containing all the cases graded as 10 (maximum) by the decision tree. The tests are (from top to bottom):

- *(a) Ratio_(10:1) >0.3318.* in agreement with Facial Thirds (ideal value of 0.333); however, although an upper bound certainly exists, the interval size is undefined;
- *(b) 0.666667 <Ratio_(16:1) <= 0.740741.* no reference value from the theories, but the test makes sense in limiting the ratio of face length to face width to a finite interval;
- *(c) Ratio_(6:7) >1.488*, in agreement with Golden Proportions' ideal value of 1.618;
- *(d) Ratio_(11:1) <= 0.3285.* in disagreement with Facial Thirds (ideal value of 0.333); however, the interval size is undefined and actual values might be very close to one third;
- *(e) Ratio_(10:1) <= 0.337-(nine), >0.337-(ten).* again, in slight disagreement with Facial Thirds (ideal value of 0.333); actually, tests (d) and (e) combined seem to state that the upper "third" in a face (see Figure 3(b)) should be slightly bigger than the mid "third" to attract the highest grades.

Since the above tests define a set of intervals, they also send the reassuring message that it is not necessary for one to have "exact" facial proportions values for achieving the highest grades of beauty. Section 3 in Fig. 12 shows the test on the root node. Such a test is in agreement with the Facial Thirds theory (ideal value of 0.333) since the sub-tree in the <= 0.3318 direction only contains grades between 1 and 7, while the opposite sub-tree contains generally higher grades, in the range 3 to 10. This shows that tests high up in the tree structure seem in substantial agreement with the aesthetics theories. This result appears striking to the Authors, given that the human referees were instructed to grade the "beauty" of a face in its entirety without any specific indication towards face proportions (and certainly had no metric cameras to measure facial features' distances and ratios).

Feature Selection and Ranking. The discussion over the classification tree reported in Figure 12 does not provide a systematic analysis of the relative importance of the various features. Therefore, we decided to use the feature selection and feature ranking algorithms in the WEKA package (Holmes, Donkin and Witten, 1994) to measure it. Feature selection algorithms apply attribute selection on input data with various search methods such as *Best-first search* and *Exhaustive search. Best-first search* evaluates a number of subsets and finds the subset with a merit of certain percentage. If the number of features selected is a subset of the original feature set, the selection criterion is good in terms of decreasing the dimensionality of the problem. In Section 2, we stated that the Facial Thirds and Golden Proportion ratios seem in good agreement. As such, they seem redundant and thus might be dropped by the selection procedure. Instead, as shown in Table 16, even with different selection methods all features were retained. Therefore, we decided to apply feature ranking on the feature set. Feature ranking algorithms rank attributes by their individual evaluations and can be used in conjunction with attribute evaluators such as *ReliefF*, *GainRatio*, *Entropy* etc. The details of the feature ranking approaches used and the ranking obtained are given in Table 17. Results show that vertical proportions seem to play a more important role in facial beauty classification than horizontal proportions. Within the vertical proportions, features from the Golden Proportions theory are the best descriptor. In general, the Golden Proportions seem to be the most significant features, and thus better descriptors than the others.

Evaluation of Classification Accuracy. A major issue in our work was how to effectively quantify the classifier accuracy given the particular ground truth. The most common figure used to describe the accuracy of a given classifier is the ratio α= *(number of correctly classified cases / total number of cases)* and the error, e, be simply defined as e = 1 - α, or:

$$e = \frac{1}{N_{img}} \sum_{i=1}^{N_{img}} e_i \begin{cases} e_i = 0 \quad if \ |t_i - a_i| = 0 \\ 1 \qquad otherwise \end{cases} \tag{5}$$

where t_i is the true class (assuming t_i = round(μ_i) from ground truth), α_i is the assigned class and N_{img} the total number of images. If the classes embed a metric as in our case (i.e.: class 3 is semantically twice as distant from class 1 than class 2), the incorrect classifications would better be weighted based on their distances from the corresponding true class. Consequently, the error rate can be redefined as:

$$e_d = \tfrac{1}{N_{img}} \sum_{i=1}^{N_{img}} d_i \ , \quad d_i = |t_i - a_i| \tag{6}$$

where d_i is the distance between the true and the assigned class.

(a)(b)

Fig. 13. Comparison between two images with similar average, but different variance values.

In our case, it must also be taken into account that the true classification deriving from ground truth is not a scalar value but rather a probability

Table 16
The feature selection approaches used and the features obtained.

Data Summary
Instances: 165
Attributes: 13
Evaluation mode: evaluate on all training data
Feature Selection
Selection Method 1:
Evaluator: weka.attributeSelection.CfsSubsetEval
Search: weka.attributeSelection.BestFirst -D 1 -N 5
Search Method: Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 61
Merit of best subset found: 0
Attribute Subset Evaluator (supervised, Class (nominal): 14 Hum_Class): CFS Subset Evaluator
Selected attributes: none
Selection Method 2:
Evaluator: weka.attributeSelection.CfsSubsetEval
Search: weka.attributeSelection.ExhaustiveSearch
Search Method: Exhaustive Search.
Start set: no attributes
Number of evaluations: 379
Merit of best subset found: 0
Attribute Subset Evaluator (supervised, Class (nominal)): CFS Subset Evaluator
Selected attributes: none

distribution. As the “true” class of each I_i image, i = 1,..., N_{img}, we assumed the rounded average of the corresponding grade histogram, t_i. Figure 13 shows the grade histograms for two limit cases from our image set. In Figure 13(a), the grade histogram is rather spread, with approximately similar values in the

Table 17
The feature ranking approaches used and the feature ranking obtained.

Data Summary
Instances: 165 Attributes: 13 Evaluation mode: evaluate on all training data
Attribute Ranking Methods
Search Method: Attribute ranking. Attribute Evaluator (supervised, Class (nominal): 14 Hum_Class): Symmetrical Uncertainty Ranking Filter
Search Method: Attribute ranking. Attribute Evaluator (supervised, Class (nominal): 14 Hum_Class): Chi-squared Ranking Filter
Search Method: Attribute ranking. Attribute Evaluator (supervised, Class (nominal): 14 Hum_Class): Gain Ratio feature evaluator
Ranking in order: **(1)** 2:4 vertical distance between pupils and tip of the chin to vertical distance between top of the face and pupils **(2)** 3:5 vertical distance between top of the face and nose to vertical distance between nostrils and tip of the chin **(3)** 6:7 vertical distance between pupils and central lip line to vertical distance between lips and tip of the chin **(4)** 5:8 vertical distance between nostrils and tip of the chin to vertical distance between pupils and nostrils **(5)** 8:9 vertical distance between pupils and nostrils to vertical distance between nostrils and central lip line **(6)** 7:9 vertical distance between lips and tip of the chin to vertical distance between nostrils and central lip line **(7)** mean_ratio **(8)** 16:1 Ratio of face width to face length **(9)** 15:16 Ratio of inter-eye distance to face width **(10)** 10:1 Ratio of vertical distance between top of the face and eyebrows to face length **(11)** 11:1 Ratio of vertical distance between eyebrows and tip of the nose to face length **(12)** 12:1 Ratio of vertical distance between tip of the nose and tip of the chin to face length **(13)** 13:14 Ratio of vertical distance between the tip of nose and lips to vertical distance between lips and tip of the chin

Table 18
The evaluation results of human and classifier classification for the testing set of 50 images.

Average of st. dev. on the test set	Average of absolute standardized distance (e_{sd})	Average of signed standardized distance (e_{ssd})
1.537	0.964	-0.085

range of 3 - 6. In Figure 13(b), instead, grades are highly concentrated around the dominant one, 5. It is evident that the distance between the true class, t_i, and the class assigned by the classifier, α_i, should be considered as errors of different weight in these two cases. Hence, in place of the simple distance, we can consider the *standardized* (or Mahalanobis) distance:

$$sd_i = \frac{|t_i - a_i|}{\sigma_i} \quad i = 1, ..., N_{img} \tag{7}$$

where σ_i is the standard deviation of the probability distribution for the I_i image. In this case, the definition of the error changes as follows:

$$e_{sd} = \frac{1}{N_{img}} \sum_{i=1}^{N_{img}} sd_i \tag{8}$$

The error redefined according to Eqs. 7-8 is an effective description of the average classifier performance over a collection of images. Table 18 provides details on the evaluation results from an experiment with the testing set of 50 images. Results obtained with the C4.5 classifier gives a value of $e = 0.964$. This result means that the standardized classifier error is on average less than the standard deviation within the class, which is inherent, proving the accuracy of classification.

Another measure of the classifier's performance can be given by the signed error, e_{ssd}, which can be obtained from Eq. 8 by simply replacing the standardized distance $sd_i = |a_i - t_i|$ by the signed distance $ssd_i = a_i - t_i$. In a balanced situation, the signed error should be close to zero since positive and negative errors of similar entity would compensate. In the experiment, we obtained a value for e_{ssd} = -0.085, thus exhibiting *undershooting*, i.e. the classifier providing a mark which is on average slightly lower than that of the human referees. However, the entity of such undershooting is very limited and the classifier can be still generally considered balanced.

5.2 Validation with the 5-fold cross-validation method

For this validation method, we divided the initial data set into five independent subsets of equal size, namely D1 to D5. Using C4.5, we then performed five iterations of training and testing. In the generic iteration i, D_i is used as the test dataset and the remaining four subsets are used for training. Such an error is in general a better estimate of the "real" error (expected value of it) than the estimate obtained from one single *holdout test* (Dietterich, 1998). The accuracy of each classifier was computed according to Eq. 8, with five error measures obtained from the various folds. In order to calculate the overall accuracy, the errors were eventually averaged. The result obtained, e_{avg}, is equal to 0.920. Such a value is very close to that of the same error measurement, e_{sd} that we obtained with *the Holdout Method* as described in the previous section and proves the stability of the accuracy value over the population.

6 Conclusions and future work

In the first part of this paper, we evaluated the extent of beauty universality by asking a diversified set of human referees to grade a collection of female facial images in terms of their facial beauty. Results obtained show that the different individuals generally provided unimodal and compact grade histograms, thus well supporting the concept that perception of beauty is universal to a certain degree.

Later, we introduced an approach to automatically measure beauty based on automated extraction of facial features and supervised classification. We presented an efficient procedure for automatically measuring facial features from face images by means of image analysis operators. For supervised classification, we used such extracted facial features and the average human grades from a set of images to train an automated classifier. The high accuracy achieved on an independent test set and from cross-validation proves that this classifier can be effectively used as an automated tool to reproduce an "average" human judgement on facial beauty. Many potential applications can benefit in the entertainment and cosmetic industries and plastic surgery. In particular, in plastic surgery this program can be used as a predictive, guiding device to decide the scope and extent of surgical corrections - given the physical constraints of each specific case. In this way, the prediction of the beauty level achievable with the surgical operation will not be based only on the surgeon's or patient's opinion, but on the estimate from a tool capable of reproducing a more enlarged human judgment. Given that the highest grades of beauty can be obtained from a range of values of facial proportions, hopefully, such a device could discourage some patients from requiring exaggerate, risky interventions. Other possible application areas can be listed as follows.

- Affective computing, to make an autonomous system (e.g. a robot) capable of human-like aesthetic judgment.
- Virtual worlds, to morph real faces under aesthetic criteria.
- The cosmetic industry, as an automated judgment tool for assessing the effectiveness of cosmetic intervention.

An important observation is that, during the poll, the human referees were asked to grade facial "beauty" in its entirety with no mention of particular regard to facial proportions. It is likely that several other features such as the color and shape of eyes, lips, nose, smiling vs. angry attitude, subject's apparent age, and others have influenced their grades. However, the fact that the automated system proved capable of reproducing the average human judgement based on facial proportions alone, gives evidence to the important role played by facial proportions in the perception of facial beauty. In the future, we plan to explore the importance of the other aforementioned factors for the extension and refining of the automated classification.

Finally, in the recent literature, there has been some attempt to also explore male facial beauty (Peseo, 2003). In his papers (Peseo, 2002), (Peseo, 2002) and (Peseo, 2003), Peseo describes the similarities and the slight differences of ratios and measurements for either gender to be considered attractive. He similarly bases his analysis on the Golden Proportions and Facial Thirds rules and adds several more ratios and criteria to them derived from other canons. Eventually, extending the analysis and classification to male facial beauty by a similar automated analysis of proportions seems feasible and can be the objective of future works. Overall, our findings and those of Peseo confirm that, in a geometrical facial biotype, certain proportions will be perceived as more attractive than others in a predictable way. However, some deviations from ideals will not necessarily compromise an optimal facial aesthetic and may explain the existence of the various kinds of "charm".

References

Aarabi, P., Hughes, D. 2001. The Automatic Measurement of Facial Beauty, IEEE International Conference on Systems, Man and Cybernetics, 4, 2644 -2647.

BBC Science - the Human Face, http://www.bbc.co.uk/science/humanbody /humanface/beauty_golden_mean.shtml (last accessed August, 2002)

Bell, A. 1997. The Definition of Beauty, Nature, October/November Issue.

Blumhof, J., Stallibrass, C., 1994. Peer Assessment, Hatfield: University of Herefordshire.

Borissavlievitch, M. The Golden Number and the Scientific Aesthetics of Architecture, A.Tiranti, London.

Bottomley, H. Maximum distance between the mode and the mean of a unimodal distribution, Leydon Close, London, 6th February 2004.

Canny, J., 1986. A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(6), 679-698.

Chen, A.C, German, C., Zaidel, D.W., 1997. Brain asymmetry and facial attractiveness: Facial beauty is not simply in the eye of the beholder. Neuropsychologia, 35, 471-476.

Cunningham, M.R., Roberts, A.R, Barbee, A.P., Druen, P.B et al., 1995. Their Ideas of Beauty are, on the Whole, the Same as Ours. Journal of Personality and Social Psychology, 68, 261-279.

Daibo, I., 1999. Suggestion From Comparison Research Of Facial Beauty, Hokusei Gakuen University, Sapporo, Japan.

Dietterich, T. G., 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation, 10(7).

Epker, N.B. and Koury, M.E., 1992. Maxillofacial Esthetics: Anthropometrics of the Maxillofacial Region. J. Oral Maxillofac. Surg., 50, 806-820.

Evans, K.M. and Kuenning, G.H. 2002. A Study of Irregularities in File-Size Distributions. International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS '02).

Farkas, Leslie G. et al., 1985. Vertical and Horizontal Proportions of the Face in Young Adult North American Caucasians. Plastic and Reconstructive Surgery, 75(3), 328-38.

Farkas, Leslie G. et al., 1987. Anthropometrics and art in the aesthetics of women's faces. Clinics in Plastic Surgery, 14(4).

Farkas L.G., 1994. Anthropometrics of the Head and Face, Second Edition, Raven Press, New York.

Graf, H.P., Chen, T., Petajan, E., Cosatto, E., 1995. Locating Faces and Facial Parts. Proc. First Int'l Workshop Automatic Face And Gesture Recognition, 41-46.

Gunes, H., Karsligil, M.Y., 2002. Measuring Female Facial Beauty by Calculating the Proportions of the Face. Proc. of ISCIS XVII Seventeenth Int. Symp. on Computer and Information Sciences, 71-75, Orlando, Florida, Oct. 2002.

Hjelmas, E., 2002. Face Detection: A Survey. Computer Vision and Image Understanding, 83, 236-274.

Holmes, G., Donkin, A., and Witten, I.H., 1994. WEKA: A machine learning workbench. Proc of Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia, 1994.

Huntley, H. E., 1970. The Divine Proportion: A Study in Mathematical Beauty, Dover Publications, New York.

Jefferson, Y., 1993. Facial Aesthetics–Presentation Of An Ideal Face, Journal of General Orthodontics, 4, 18-23.

Landau, T., 1989. About Faces, Bantam Doubleday Dell Publishing Group

Inc., New York.

Langlois, J.H., Roggman, L.A., 1990. Attractive Faces Are Only Average. Psychological Science, 1, 115-121.

Larrabee, W., 1997. Facial Beauty: Myth Or Reality?. Archives of Otolaryngology-Head and Neck Surgery, 123, 571-572.

Marquardt Beauty Analysis, 2002. http://www.beautyanalysis.com/Index2_mba.htm (last accessed 12/07/2004).

Mealey, L., Bridgstock, R., Townsend, G.C. 1999, Symmetry and Perceived Facial Attractiveness: A Monozygotic Co-Twin Comparison.

Michalski, R. S., Carbonell, J. G., Mitchell, T. M., 1984. (Eds.), Machine Learning - An Artificial Intelligence Approach. Springer-Verlag, Berlin.

Michiels, G., Sather A.H. 1994. Determinants of Facial Attractiveness in a Sample of White Women, Int. Journal of Adult Orthodontics and Orthognathic Surgery, 9, 95-103.

Parris, C., Robinson, J. Jr. 1999. The Bold And The Beautiful (According To Plastic Surgeons), Tyler Street Christian Academy and University of Texas Southwestern Medical Center, Dallas, Texas, http://www.swmed.edu/stars/resources/stock99/parris.html (last accessed 12/07/2004).

Perrett, D.I., May, K.A., Yoshikawa, S. 1994. Facial Shape and Judgments of Female Attractiveness, Nature, 239-242.

Peseo, G., 2002 (a). The “Beauty” of Homo sapiens: standard canons, ethnical, geometrical and morphological facial biotypes (part one), Virtual Journal of Orthodonics, Vol. 4, No. 4.

Peseo, G., 2002 (b). The “Beauty” of Homo sapiens: standard canons, ethnical, geometrical and morphological facial biotypes (part two), Virtual Journal of Orthodonics, Vol. 5, No. 1.

Peseo, G., 2003. The “Beauty” of Homo sapiens: standard canons, ethnical, geometrical and morphological facial biotypes (part three), Virtual Journal of Orthodonics, Vol. 5, No. 2, ISSN - 1128-6547.

Quinlan, J. R., 1993. C4.5: Programs For Machine Learning, Morgan Kaufman Publishers, San Mateo, California.

Rein-Lien Hsu, Abdel-Mottaleb, M., Jain, A.K., 2002. Face Detection in Color Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (5), 696 -706.

Ricketts, M.D., 1982. Divine Proportions in Facial Aesthetics. Clinics in Plastic Surgery, 9(4).

Two-Sample *t*-Test for Equal Means, NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/ (last accessed 31/05/2006)

The Golden Proportion, http://www.goldenmeangauge.co.uk/golden.htm, http://www.beautyanalysis.com/mba_phithekeytobeauty_page.htm (last accessed 12/07/2004)

Yang, M., Kriegman, D.J., Ahuja, N., 2002. Detecting Faces In Images: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(1).

Yellin, S., 1997. Aesthetics for the Next Millennium. Facial Plastic Surgeons Monographs, 13(4), 231-239.

Yow, K.C., Cipolla, R., 1995, Finding Initial Estimates of Human Face Location, Proc.2nd Asian Conf. on Comp. Vision, 3, 514-518.

Yow, K.C., Cipolla, R., 1996. A Probabilistic Framework For Perceptual Grouping of Features for Human Face Detection, Proc. Second Int'l Conf. Automatic Face And Gesture Recognition, 16-21.

Yow, K.C., Cipolla, R., 1997. Feature-Based Human Face Detection. Image and Vision Computing, 15(9), 713-735.