CALIBRATION, VALIDATION, AND EVALUATION OF SCANNING SYSTEMS:

Anthropometric Imaging System Repeatability

Luci Ann P. Kohn
and
James M. Cheverud

Department of Anatomy and Neurobiology
Box 8108, 660 S. Euclid Avenue
Washington University School of Medicine
St. Louis, Missouri 63110
Telephone: (314) 362-4189
Fax: (314) 362-3446
E-Mail:
KOHN_L@WUMS2.WUSTL.EDU (LAPK)
CHEVERUD_J@WUMS2.WUSTL.EDU (JMC)

Introduction

Human body surface dimensions are useful to researchers in a number of disciplines, including anthropometrists who characterize the morphology of a population, engineers who develop clothing or protective devices, plastic and reconstructive surgeons who diagnose and treat abnormal morphology, and forensic scientists who attempt to reconstruct facial dimensions from cranial material.

These disciplines share the necessity of gathering accurate, highly repeatable measurements of homologous anatomical structures, either landmarks or surfaces. Traditional anthropometric studies are based on anatomical locations (or landmarks) and dimensions defined by Hrdlicka (1920). Anthropometers, calipers, and tape measures have been used to collect these dimensions. The traditional measures have been used to characterize and compare a wide variety of human populations. The standard list of measurements has also been augmented for use in medicine (Farkas, 1981) and human engineering. Traditional anthropometric techniques have been limited to linear distances of unknown configuration in three-dimensional space.

Several recent studies have introduced new methods for the measurement of three-dimensional surface landmark coordinates and the use of these coordinates to reconstruct surfaces and volumes. Burke and coworkers (Burke and Beard, 1979; Burke and Hughes-Lawson, 1988) used stereophotogrammetry to collect three-dimensional locations on the human face. Cheverud et al. (1990a, 1990b) and

Donelson and Gordon (1991) report variation and covariation among measurements based on three-dimensional coordinates of head landmarks obtained from over 9,000 US Army personnel with an automated measurement device (Annis and Gordon, 1988). These methods are accurate in the collection of three-dimensional data (Annis and Gordon, 1988).

Our position is that it is necessary to evaluate the precision and repeatability of all new surface imaging devices, and to compare the new device to a simpler and already validated device for collecting morphological data. We believe that it is necessary to determine the repeatability and precision of the methods used in all aspects of data collection.

Body Imaging System: Requirements and Validation

There are currently a number of research groups developing new devices for capturing three-dimensional images of body surfaces. Several factors must be considered in the validation and testing of an imaging system for anthropometric measurement:

1) The digital images produced must be as clear and accurate as possible, allowing visual identification of contrasting colors and depths by an observer or by electronic means. That is, error in detecting marked landmarks in digitized images should be evaluated.
2) Repeated images of the same individual should be nearly identical. That is, error in producing a digital image from a living subject should be evaluated.
3) The locations of the anatomical structures chosen for data collection must be reliably identifiable by a trained observer. That is, error in locating and marking landmarks by the anthropometrist should be evaluated.

Tests of new imaging devices need to deal with these three issues in terms of precision and repeatability. Precision is the average absolute difference between repeated measures of the same individual. Repeatability is the precision of the measurement relative to the differences among individuals contrasted in any given study. It is measured as the proportion of total variance due to differences between individuals using a repeated measures analysis of variance design. One minus the repeatability is the proportional error due to measurement inaccuracy. The level of repeatability and the imprecision which can be tolerated in any particular study will vary with the magnitude of the contrast being made. For example, measurements which are precise enough for a comparison of infants with adolescents may not be precise enough for studying variation among adolescents. Usually, we judge repeatability within a single homogeneous population since this places the most stringent constraint on error levels.

While the concepts described above determine the replicability of measurements

taken using a device, these measurements may still be inaccurate or biased. The potential bias of the measurement system can be judged by validation studies in which measurements taken from some simpler, already validated system are compared to those obtained by the new device.

We outline one design for testing and validating electronic imaging systems. This procedure, depicted in Figure 1, allows a researcher to test whether structures can be reliably located on the images produced by the imaging system, whether an imaging system can produce repeatable images, and whether structures can be reliably located on individuals. The structures to be tested can be anatomical landmarks, surfaces or linear dimensions; however, for ease of reference our discussion will focus on the analysis of anatomical landmarks.

Data Collection

Prior to recruitment of volunteers, landmarks or surfaces are identified for study. For example, if the imaging device is to be used in plastic surgery applications, facial contours or landmarks may be identified which are commonly used in characterizing normal and abnormal facial morphology. This would include landmarks or surfaces which would be changed by plastic surgery (e.g., regions around the eyes, the cheeks, or the nose), and those which may be expected to remain unchanged by surgical intervention (e.g., the ears).

The number of volunteers necessary for a validation study is dependent on whether the imaging device will be used to characterize a single population (within population) or differences between two or more populations (between population). For example, approximately 10 volunteers should be recruited for the repeatability study within a single population. This sample size is sufficient to provide a general level of precision and repeatability. It allows us to identify which component, or components, of the measurement process is the major source of error so that this component can be targeted for further development and improvement.

Following the scheme presented in Figure 1, each volunteer is scheduled for two separate measurement sessions. At each measurement session, landmarks are first identified and marked. Differences in measurements taken on these two occasions are due to error in landmark localization and marking. Two images of each marked individual are collected at each measurement session. Differences between these two images are due to error in digital recording by the device being tested. Finally, the three-dimensional landmark coordinates are recorded twice from each digital image. Differences in measurements taken at separate times on the same image are due to recording error. Thus we can separate error due to (1) locating landmarks on subjects; (2) the imaging device itself; and (3) the digitizing of landmarks from the images. With this design, the data for each individual in the repeatability study consists of eight sets of landmark locations.

Precision and Repeatability

The statistical analysis of the repeated measures allows us to estimate the precision and repeatability of the measurement device. Precision of three-dimensional coordinates can be measured by superimposing repeated images in such a way that the squared distance between homologous landmarks is minimized. This registration procedure is called Procrustes rotation (Goodall and Bose, 1987).

First, the two separate digitizations of a single image are compared by deriving the distance between homologous landmark positions after Procrustes rotation. These differences are averaged over images, measurement sessions, and individuals to provide a measure of error due to the digitizing process. Error at this level is the result of interaction between the device and the operator.

Second, the two digitizations of each image are averaged (after Procrustes rotation). These averages represent each of the two images obtained at each measurement session. The image averages are then compared to one another after Procrustes rotation. The distance between homologous landmarks is calculated for each pair of images taken at a single measurement session and then averaged over measurement sessions and individuals to provide a measure of error due to the imaging device.

Third, the two image averages obtained at each measurement session are averaged (after Procrustes rotation). These new averages represent the separate landmark markings done at each measurement session. The marking averages are then compared to one another after Procrustes rotation. The distance between homologous landmarks is calculated for each pair of coordinates from a single individual and then averaged over individuals to provide a measure of error due to marking by the anthropometrist.

With these three sets of comparisons we can measure imprecision due to digitizing from the images, marking the images, and marking the landmarks on a living subject.

Repeatability can also be measured with this design. First, a Procrustes rotation is performed using all coordinate sets (8 per individual) in a single analysis. The X, Y, and Z coordinates are then used as dependent variables in a nested analysis of variance with three factors, individual, measurement session nested within individual, and image nested within measurement session and individual. The separate digitizations of each image serve as the residual term. The variance due to each of these random effects is derived from the analysis of variance. The proportion of the total variance due to individual differences is the repeatability. The proportion of the total variance due to measurement session represents error due to marking. The proportion of the total variance due to image represents error due to the imaging device. Finally, the proportion of the total variance due to the residual represents the error due to digitization of the images.

The procedures described above for three-dimensional coordinate data can also be applied to other kinds of data derived from anthropometric images. One can derive linear distances and angles either directly from the subjects or from the digitized images. Alternatively, surfaces can be compared after registration using some measure of difference between the repeated surfaces, such as the volume separating the surfaces. This volume would be treated in the same way as the distance between repeated landmarks in the discussion above.

Validation consists of comparing the results from the new imaging device to results obtained from other, previously validated, techniques. These comparisons would be done using Procrustes techniques, as described above. The Procrustes-derived average coordinate locations for each individual obtained with the new device are compared, using Procrustes rotation, to the locations derived for that same individual using the valid technique. Distances between homologous landmarks indicate error in the new device.

Example: Validation of the Cencit 3D Facial Surface Scanner

We are involved in a validation study of the Cencit 3D digitizing system, a facial surface scanner. Our objective is to test whether the Cencit system is suitable for applications in anthropometry and medical imaging of the human face. Our study was designed to test whether the Cencit system produces repeatable facial images and whether landmarks could be accurately located on the images produced by the Cencit facial scanner.

We recruited 10 normal males and 10 normal females from the Washington University and University of Missouri - St. Louis community. All of these volunteers were of Northern European descent. In addition, one female of African descent and one male of Asian descent were measured. These two individuals were not included in the main portion of the validation study, but their data were compared to that of the Northern European volunteers to see if their landmarks could be identified reliably.

Twenty-seven facial landmarks (Table 1) were identified on each individual, and the landmark locations were marked with a 1/4" non-shiny black adhesive dot. The landmarks chosen were used in a previous anthropometric study of facial morphology (Cheverud et al., 1990a, 1990b) and their locations had been found to be highly reproducible (Annis and Gordon, 1988). We followed the measurement scheme outlined above. On each of two measurement sessions, landmarks were marked on each participant, and two Cencit facial scanner images were recorded. Each Cencit scanner image was digitized twice, yielding 8 sets of three-dimensional coordinates for each individual. All landmark location and digitizing were performed by one individual to eliminate interobserver error. The same procedure was repeated with the Polhemus 3Space digitizer which has been subjected to

previous validation studies (Hildebolt and Vannier, 1988). Only data from the females were used in the analyses presented below.

Results of Cencit Facial Scanner Analysis

The results of the analysis of precision are summarized in Table 2. For the Cencit facial surface scanner, the minimum and maximum distances between homologous landmarks at the level of digitizing (0.07 - 0.15 cm), scanning (0.07 - 0.22 cm), landmark marking (0.17 - 0.56 cm) are all less than the diameter of the black dots used to mark the landmarks. Homologous landmarks measured by the Polhemus 3Space digitizer have comparable minimum and maximum distances (scanning: 0.10 - 0.20 cm; landmark location: 0.18 - 0.70 cm). The precision of the Cencit surface scanner is comparable to that of the Polhemus 3Space digitizer. The greatest amount of error is in the marking of the landmarks. However, the magnitude of the imprecision is smaller than the diameter of the black dots used to mark the landmarks. The size of the black dots used was determined by requirements of the Cencit scanner. We can expect greater precision when smaller markers can be used

The average proportion of the total variance explained by digitizing, scanning, landmark location, and individual differences (repeatability) is presented in Table 3. In the location of the X, Y, and Z coordinates from the Cencit facial scanner, an average of 6% of the total variation is due to differences in digitizing, 4% of the total variation is due to differences in scanning, and 31% of the variation is due to differences in locating the landmarks in the two measurement sessions. The repeatability of the X, Y, and Z landmark coordinates measured with the Cencit surface scanner averages 59%. For X, Y, and Z landmark coordinates measured with the Polhemus 3Space digitizer, an average of 8% of the variation is due to differences in scanning; 23% of the variation is due to differences in landmark location between the two measurement sessions. The repeatability of the X, Y, and Z landmark coordinates measured with the Polhemus 3Space digitizer averages 69%. Measurements of bilateral distances (eg., bizygomatic breadth, biectocanthus breadth) are more highly repeatable, with 76% repeatability using the Cencit facial surface scanner, and 84% repeatability using the Polhemus 3Space digitizer. The proportions of variance explained by digitizing, scanning, and landmark location are roughly comparable between the two systems.

The precision and repeatability of the Cencit surface scanner and the Polhemus 3Space digitizer are similar. The proportion of the variance due to digitizing and scanning is low. A large proportion of the variance within individuals is due to error in landmark location. This is human error and can be reduced by training of the anthropometrist. A reduction of this human error will increase the repeatability. The repeatability of bilateral distances is greater than repeatability of the X, Y, and Z landmark coordinate locations. The imprecision of locating the landmarks is generally much less than 0.64 cm, and we are therefore explaining the repeatability of finding a landmark within a very small region. Small

measurements can be expected to be less repeatable than large measurements. Large measurements can be expected to be measured with greater repeatability because the errors are distributed across a greater distance.

Evaluation of the Cencit Facial Scanner

The Cencit surface scanner performs comparably to the Polhemus 3Space digitizer. The Cencit surface scanner accurately images facial landmarks. A large amount of data can be collected from electronically stored facial images. Images can be collected and processed at a researcher's convenience, and stored images may be useful to a number of researchers for a variety of purposes. Such a system will be useful in studies of landmarks, surfaces, and volumes and in comparisons within or between populations. The majority of the error in our study was found to be in the location of the landmarks, i.e., human error. This source of error will be common to all measurement devices and can be reduced by training and judicious choice of landmarks.

Suggestions for Future Improvement

A number of problems arose in our validation study of the Cencit Facial Scanner, representing areas in which additional effort is needed before this electronic imaging system is fully suitable for an anthropometric application.

Not all desired landmarks were easily visible on the scanned images. Landmarks which were not visible were regarded as missing. We identified landmarks on all areas of the face and ears; however, ears were rarely imaged sufficiently to identify the marked landmarks. In addition, landmarks on the margins of the face were often unobservable due to variations in lighting. Landmarks in the center of the face were the most clearly visible because the system's cameras were concentrated on this region. These shortcomings can be overcome with changes in lighting, camera location, and data processing. Future developments may also enable landmark locations and their defining surfaces to be identified electronically, perhaps by locating landmarks of a particular color density. This will reduce error and time spent in data acquisition.

References

Annis, J. F. and Gordon, C. C. (1988) Development and validation of an automated headboard device for measurement of three-dimensional coordinates of the head and face. Technical Report NATICK/TR-88/048. U. S. Army Natick Research, Development and Engineering Center, Natick, Massachusetts.

Burke, P. H. and Beard, L. F. H. (1979) Growth of soft tissues of the face in

adolescence. Brit. Dent. J. 146:239-246.

Burke, P. H. and Hughes-Lawson, C. A. (1988) The adolescent growth spurt in the soft tissues of the face. Ann. Hum. Biol. 15:253-262.

Cheverud, J., C. C. Gordon, R. Walker, C. Jaquish, L. Kohn, A. Moore and N. Yamashita. 1990a. Anthropometry survey of U. S. Army Personnel: Correlations and regression equations. Technical Report NATICK/TR-90/022. U. S. Army Natick Research, Development and Engineering Center, Natick, Massachusetts.

Cheverud, J., C. C. Gordon, R. Walker, C. Jaquish, L. Kohn, A. Moore and N. Yamashita. 1990b. Anthropometry survey of U. S. Army Personnel: Bivariate Frequency Statistics. Technical Report NATICK/TR-90/023. U. S. Army Natick Research, Development and Engineering Center, Natick, Massachusetts.

Donelson, S. M. and Gordon, C. C. (1991) 1988 Anthropometric Survey of U. S. Army Personnel: Pilot Summary Statistics. Technical Report NATICK/TR-91/040. U. S. Army Natick Research, Development and Engineering Center, Natick, Massachusetts.

Farkas, L. G. (1981). Anthropometry of the Head and Face in Medicine. New York:Elsevier.

Hildebolt, C. F. and Vannier, M. W. (1988) Three-dimensional measurement accuracy of skull surface landmarks. Am. J. Phys. Anthrop. 76:497-503.

Hrdlicka, A. (1920) Anthropometry. Philadelphia:Wistar Institute.

Goodall, C. and Bose, A. (1987) Models and Procrustes methods 19th Symposium of the Interface Between Computer Science and Statistics, pp. 86-92.

Figure 1. Scheme for validation study of Cencit 3D surface scanner. For each individual, I , identification of landmarks (measurement session), facial surface images and digitizing (d_i) of identified landmarks are each repeated twice to estimate the amount of variance due to measurement error as compared to variance due to individual differences. The measurement error can then be divided into error in identifying landmarks, error in producing an image from a living subject, and error in digitally recording landmark locations.

Individual I							
Measurement Session 1				Measurement Session 2			
Image 1		Image 2		Image 1		Image 2	
d_{11}	d_{12}	d_{21}	d_{22}	d_{11}	d_{12}	d_{21}	d_{22}

Table 1. Landmarks included in anthropometric analysis. Landmarks 1 through 11 are bilateral landmarks and were measured on the right and left sides. Landmarks 12 through 16 are unilateral landmarks located in the midsagittal plane.

1. Alare (Right and Left)
2. Cheilion (Right and Left)
3. Ear Top (Right and Left)
4. Ectocanthus (Right and Left)
5. Frontotemporale (Right and Left)
6. Gonion (Right and Left)
7. Infraorbitale (Right and Left)
8. Otobasion Superior (Right and Left)
9. Tragion (Right and Left)
10. Zygion (Right and Left)
11. Zygofrontale (Right and Left)
12. Chin
13. Crinion
14. Promenton
15. Pronasale
16. Sellion

Table 2. Average precision (in cm) of landmark locations measured from the Cencit surface facial scanner and the Polhemus 3Space digitizer. This is average minimum and maximum distances between homologous due to differences in digitizing landmarks, scanning the image, marking the landmarks, and differences between individuals. Note that the dots used to mark the landmarks are 0.64 cm in diameter.

	Average Minimum	Average Maximum
Cencit		
Digitize	0.07	0.15
Scan	0.07	0.22
Marking	0.17	0.56
ID	0.05	1.30
3Space		
Scan	0.10	0.20
Marking	0.18	0.70
ID	0.02	1.60

Table 3. Average proportion of total variation in location of X, Y, and Z, coordinates and bilateral distances explained by digitizing, scanning, marking landmarks, and individual (repeatability) for the Cencit surface scanner and the Polhemus 3Space digitizer.

	X, Y, Z axis Location	Bilateral Distances
Cencit		
Digitize	6%	4%
Scan	4%	4%
Marking	31%	16%
ID	59%	76%
3Space		
Scan	8%	11%
Marking	23%	5%
ID	69%	84%