ELSEVIER

Journal of Neuroscience Methods 115 (2002) 137–143

JOURNAL OF NEUROSCIENCE METHODS

www.elsevier.com/locate/jneumeth

# A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies

Ruben C. Gur[a,*], Radim Sara[b,c], Michiel Hagendoorn[a], Oren Marom[a], Paul Hughett[a], Larry Macy[a], Travis Turner[a], Ruzena Bajcsy[b], Aaron Posner[d], Raquel E. Gur[a]

[a] *Brain Behavior Laboratory, Neuropsychiatry Division, Department of Psychiatry, Philadelphia, PA 19104, USA*
[b] *The GRASP Laboratory, Department of Computer Science, University of Pennsylvania, Philadelphia, PA 19104, USA*
[c] *Center for Machine Perception, Department of Cybernetics, Czech Technical University, Prague Czech Republic*
[d] *The Arden Theatre Company, Philadelphia PA 19106, USA*

Received 18 June 2001; received in revised form 21 December 2001; accepted 24 December 2001

## Abstract

Facial expressions of emotion are increasingly being used in neuroscience as probes for functional imaging and as stimuli for studying hemispheric specialization for face and emotion processing. Available facial stimuli are 2-dimensional and therefore, their orientation is fixed and poorly suited for examining asymmetries, they are often obtained under poorly specified conditions, usually posed, lack ethnic diversity, and are of restricted age range. We describe a method for accurately acquiring and reconstructing the geometry of the human face and for display of this reconstruction in a 3-dimensional format. We applied the method in a sample of 70 actors and 69 actresses expressing happiness, sadness, anger, fear and disgust, as well as neutral expressions. Each emotion was expressed under three levels of intensity and under both posed and evoked conditions. Resulting images are of high technical quality and are accurately identified by raters. The stimuli can be downloaded in digital form as ‘movies’ where angle and orientation can be manipulated for inclusion in functional imaging probes or in tests that can be administered as measures of individual differences in facial emotion processing. The database of emotional expressions can also be used as a standard for comparison with clinical populations. © 2002 Published by Elsevier Science B.V.

*Keywords:* Emotion; 3-Dimensional faces; Polynocular stereo; Surface reconstruction; Computer vision

## 1. Introduction

Neuroscience investigations increasingly use facial expressions of emotions as probes for functional imaging (Adolphs et al., 1996; Breiter et al., 1996; Hyman, 1998; Morris et al., 1996, 1998) and in tasks that can be correlated with neuroanatomic measures and effects of brain disorders (Adolphs et al., 1994; Blonder et al., 1991; Borod, 1993; Kohler et al., 2000). Studies have applied stimuli of varying quality obtained under differing conditions, most including posed emotions. Facial stimuli are typically of a restricted ethnicity and age range. Furthermore, they are 2-dimensional photographs, where facial orientation is either poorly controlled or artificially made at straight angle. Such 2-dimensional stimuli are not amenable for manipulations of angle and orientation, and raise methodological concerns when applied to examination of facial asymmetries which could be related to hemispheric specialization (Sackeim et al., 1978; see review in Borod et al., 1997).

We describe a method for obtaining digitized high-quality 3-dimensional photographs of facial expressions and its implementation under standardized conditions in a sample of 139 actors (70 male, 69 female) of diverse ethnicity and age. The method has generated a digital database of facial expressions of happiness, sadness, anger, fear, and disgust, under both posed and evoked conditions, each at three levels of intensity, and neutral expressions. The validity of the emotions expressed was

[*] Corresponding author. Tel.: +1-215-615-3604; fax: +1-215-662-7903.
*E-mail address:* gur@bbl.psycha.upenn.edu (R.C. Gur).

PII: S0165-0270(02)00006-7

established in a sample of raters. The stimuli are available for downloading from the internet as pictures or 3-dimensional movies.

## 2. Method

### 2.1. Image acquisition methodology

We developed a protocol to accurately acquire and reconstruct the geometry of the human face and display this reconstruction in a 3-dimensional format. Such methods of 3D reconstruction from multiple stereo camera setups are called polynocular stereo. Image capture is achieved with four Pulnix TM9701 digital video cameras, and a Nikon N90 35 mm SLR camera (with Kodak Ektachrome 320T film, exposed at 1/60 s and *f*/5.6) for color rendering is centered among them. The digital cameras are mounted on a custom-made aluminum frame and attached to a tripod with magnetic 'feet.' The SLR is positioned on a second tripod at the caudal aspect of the frame with the same rostral orientation as the digital cameras. The floor supports a 250 kg, steel plate to which the magnetic feet of both tripods attach. This effectively stabilizes the setup. Image capture is triggered by a computer signal with all five cameras beginning image acquisition simultaneously. We use large-screen flood lights giving soft (diffuse) illumination. This is important for resulting texture fidelity.

The images are transferred serially from their buffer in the digital cameras through a Bitflow multiplexer onto the computer, using a software-controlled hardware interface (Vision One, Bozeman, MT; Fig. 1). The color image from the SLR is scanned into the computer off-line.

Once all five temporally corresponding images are in digital form, they are run through an algorithm that coalesces the 2D information into a 3D model. The algorithm consists of five successive stages (Bajcsy et al., 1998; McKendall et al., 1997; Sara, 2000). First, pointwise image correspondences are found by a stereoscopic matching algorithm for each of the six possible image pairs. The correspondence quality is assessed by a normalized cross-correlation similarity measure computed over a small $5 \times 5$ image window. Stable Matching Algorithm is used to select correct matches (Sara, 2001). The matching procedure is fully automatic but is prone to mismatches. They occur due to anisotropy of surface reflectance (if one camera sees a specular highlight at a certain surface point and the other does not, the matching procedure may get confused) or due to poor signal-to-noise ratio (points of low-contrast or vanishing texture may become mutually indistinguishable).

Most mismatches from the first stage are removed in the second stage of correspondence verification: each point correspondence between images A, B predicts four more correspondences in pairs A–C, A–D, B–C, B–D. Each pairwise correspondence is, therefore, validated in two independent images. Again, stable matching is used (see Sara, 2000, for details).

In Stage 3 all validated correspondences are used to reconstruct the point-cloud model in 3D space. The point-cloud model lacks connectivity. The task of Stage 4 is to reconstruct a continuous surface from the point model while rejecting the remaining outliers. The reconstruction procedure performs two steps. Probability density function is first estimated from the point cloud. The density captures the likelihood that a particular spatial region corresponds to a 3D surface. Second, a maximum-probability surface is extracted by

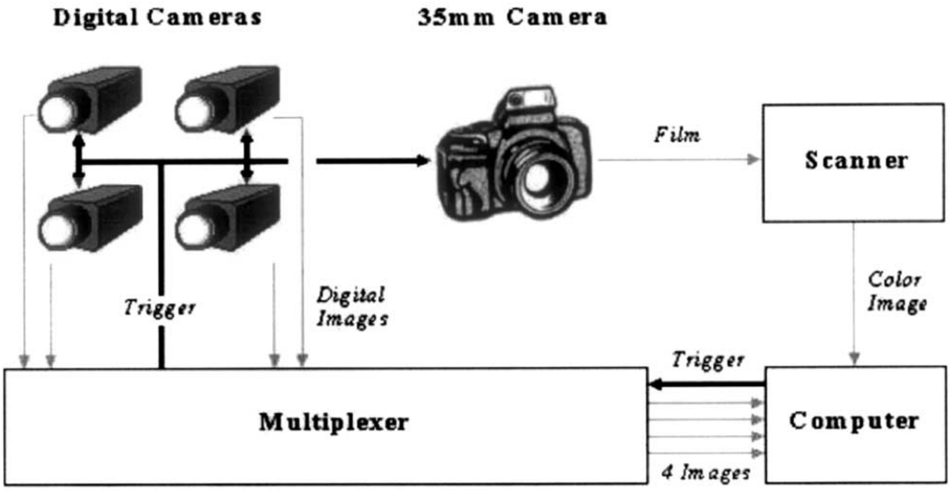

Fig. 1. Five cameras with synchronous image capture are computer controlled via multiplexer. The four digital cameras send images through the multiplexer (the stereo set) and the photographic camera film is scanned to produce high-quality color images (texture images). All image data are processed off-line to produce final 3D models.

means of a surface tracing algorithm (Sara and Bajcsy, 1998). Stage 4 results in a triangulated 3D geometric model.

The last stage of 3D model reconstruction maps texture onto the model (Fig. 4). This is done by reprojecting each control point of the triangular mesh to the texture image. Each control point thus receives five coordinates: three spatial ($x$, $y$, $z$) and two textural ($u$, $v$). From this information any standard VRML viewer can render arbitrary views of the model on the computer screen.

The quality of the 3D reconstruction depends on temporal resolution, lighting, and optical distortion, but the most important determinant is spatial accuracy. We, therefore, improve matching accuracy by projecting a random noise pattern that is IR filtered so as to appear on the digital (Fig. 2), but not the analog images (Fig. 3). The random dot pattern is projected onto the faces using a pair of Kodak Ektagraphic III AMT slide projectors, which have been modified by removing the heat absorbing glass and placing Wratten type 89B filters over the lenses; these modifications mean that the projected pattern is visible only in the near infrared range (roughly 700–1000 nm) and is invisible to the eye and to the film camera used to capture the color images. To improve visibility of the projected pattern in the digital cameras, their lenses are also equipped with additional Wratten 89B filters; this setup exploits the fact that the silicon CCDs are sensitive in the near infrared as well as in the visible light (roughly 400–700 nm). The pattern is copied onto a 2 × 2 in. of Kodak technical panchromatic film and placed between two glass mounts for use in the projector. The images of the dot pattern form Julesz random dot stereograms. The random dot size and spacing matches the Nyquist frequency of the digital cameras in order to obtain the best possible geometric resolution.

Since the slides from the photographic camera are to be scanned, we need to correct for inaccuracies of the scanning process. Wall fiducials (Figs. 2–4) are used to co-register each scanned slide with the images of the calibration target. This registration eliminates small shifts and rotations that occur due to imprecise scanning mechanics. The correction improves the quality of texture mapping.

Our current accuracy of facial reconstruction is about 0.2 mm rms on well textured surfaces sloping less than 45°. The model obtained from the five-stage surface reconstruction is a fully 3-dimensional entity embedded in a Cartesian metric coordinate system. The model conforms to the VRML V.2 specification (VRML, 2001) and consists of a dense triangular mesh approximating the surface. The mesh control point spacing slightly varies around 4 mm. A single VRML model including texture occupies about 0.6 MB of disk space.

The VRML model is a primary output format. Secondary is the MPEG movie format created automatically from the primary format by means of a scriptable geometric viewer (Phillips, 1993). The MPEG movie shows the head rotating in the transversal plane. Each movie consists of 33 full-quality 480 × 352 frames and

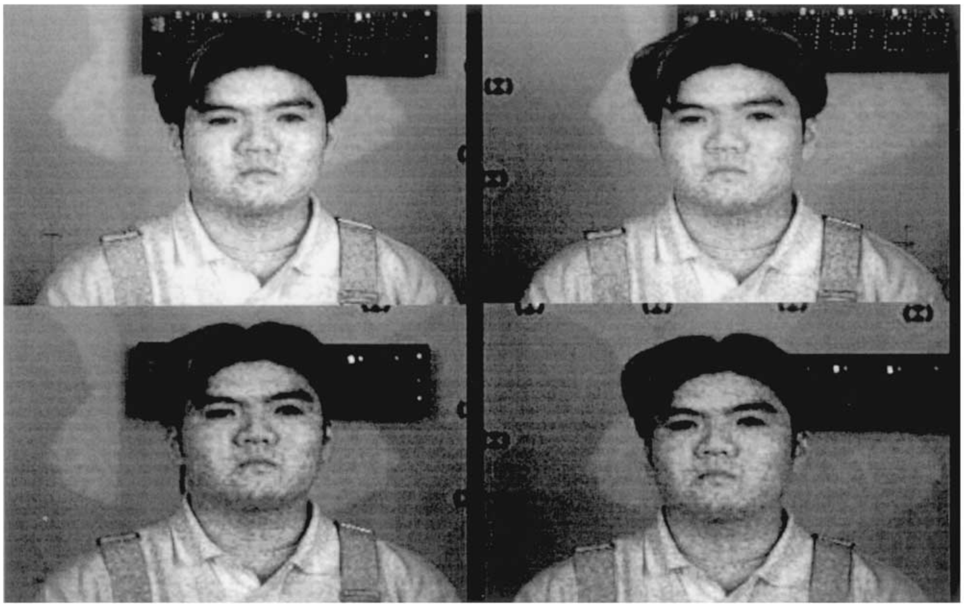

Fig. 2. Example of the images captured by the four digital cameras. The digital clock visible in the background was used as a backup method of relating the film and digital images to each other and to the handwritten session notes. The clock was obtained from Jameco (Belmont, CA, part #JE725AEA) and the original red LEDs were replaced by type XC554 R LEDs so that they would be visible to both the film and digital cameras. The noise pattern on the face enables accurate 3D reconstruction.

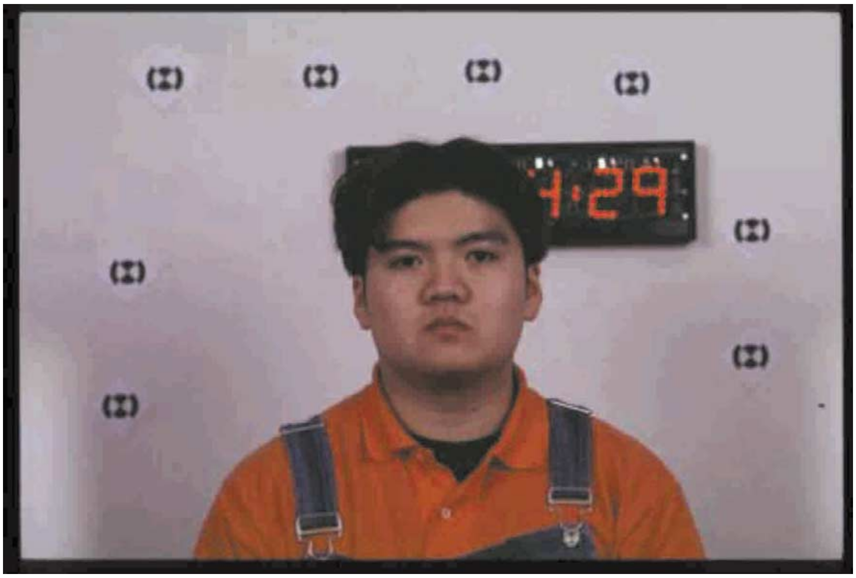

Fig. 3. The texture image used for color rendering. The wall fiducials are used to co-register the scanned slide with the images of camera calibration target. The registration eliminates small geometric errors due to imprecise scanning mechanics.

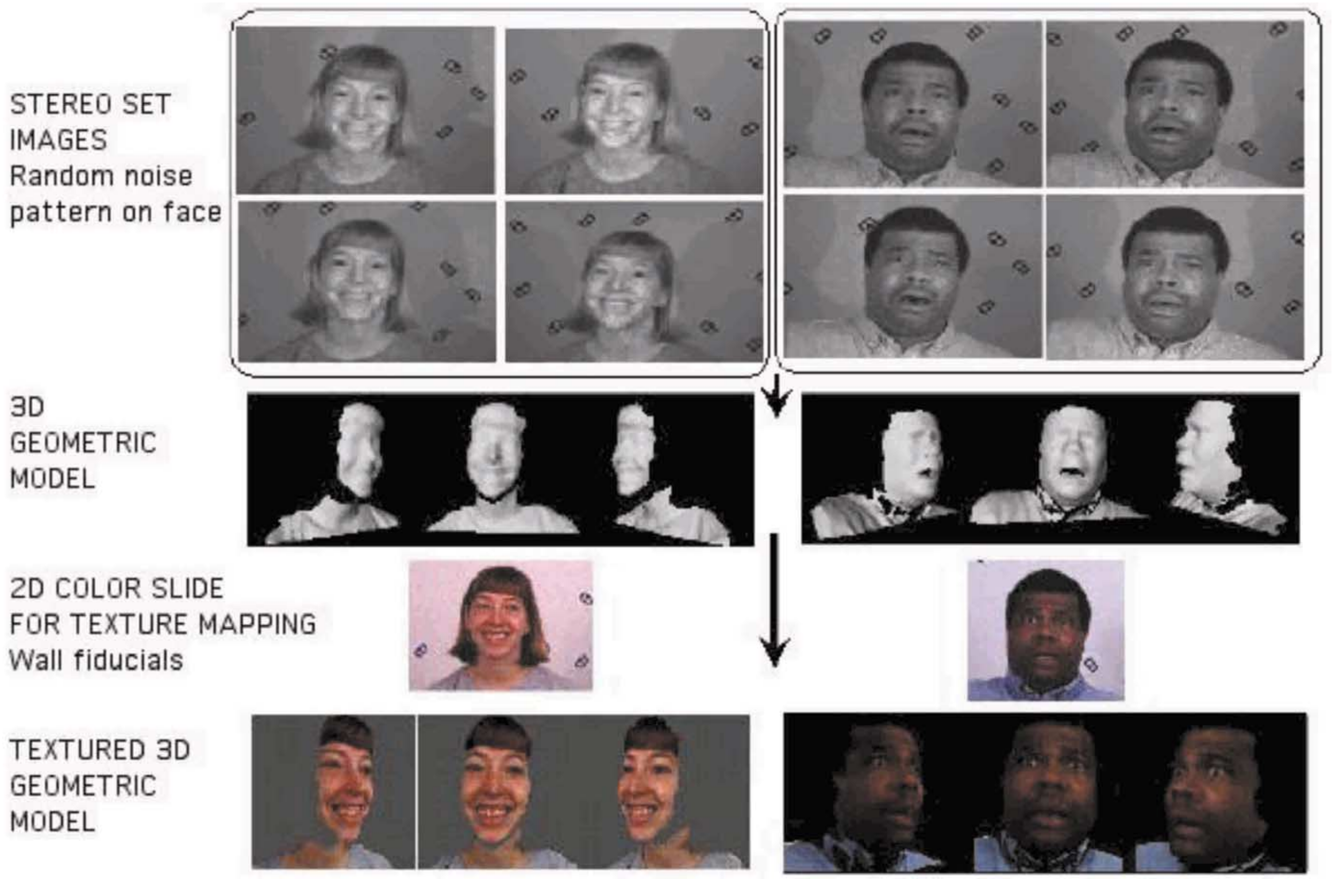
STEREO SET IMAGES
Random noise pattern on face
3D GEOMETRIC MODEL
2D COLOR SLIDE FOR TEXTURE MAPPING
Wall fiducials
TEXTURED 3D GEOMETRIC MODEL

Fig. 4. The steps in the process of 3D geometric reconstruction with texture map. Images from the stereo set are processed by stereo matching and surface reconstruction algorithms to produce a 3D geometric model of the object. The color texture image is then blended with the model using texture-mapping technique. The final model is a three-dimensional virtual entity, which can be viewed from an arbitrary viewpoint.

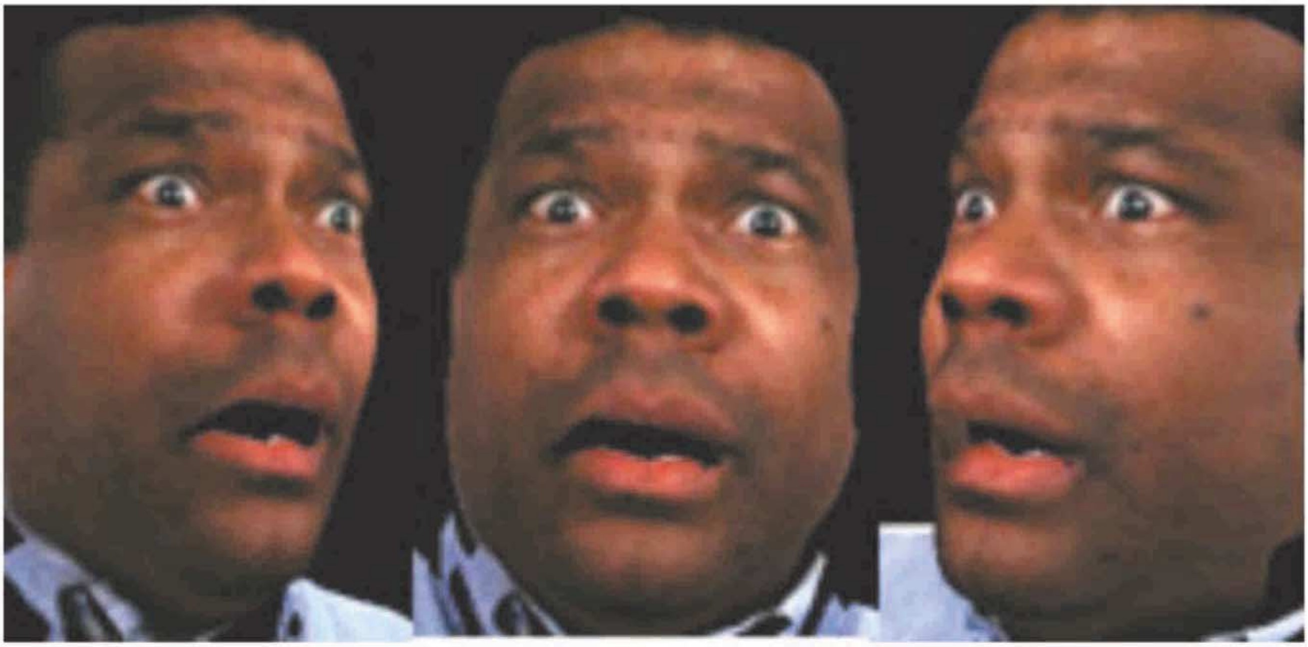

Fig. 5

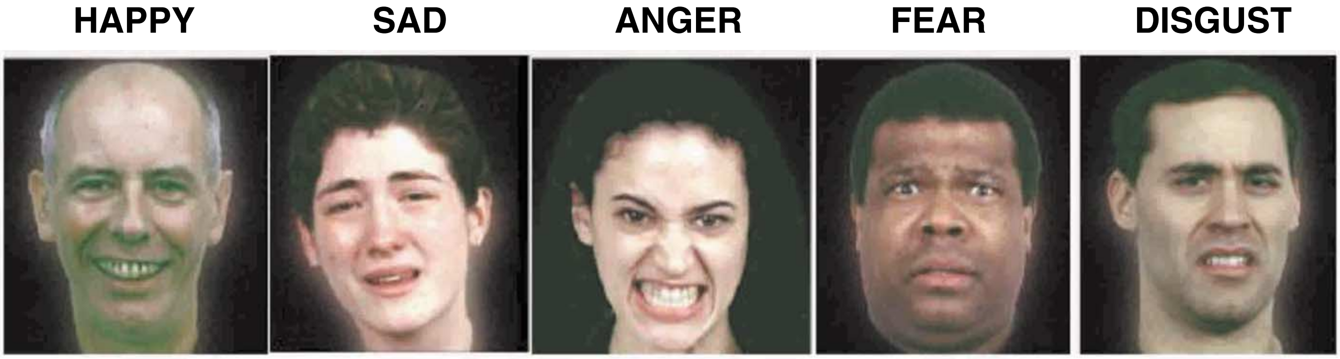

Fig. 6

Fig. 5. An example of the reconstruction with a fearful expression, illustrating three projections.

Fig. 6. Examples of expressions for the fie emotions. Note the uniformity of results across age and ethnicity.

occupies also about 0.6 MB. Thus, single frames can be obtained at any desired angle (Fig. 5)

In order to reconstruct a metric model, all cameras must be calibrated. Each camera is described by 11 parameters: focal length, aspect ratio, image grid skew, the image center (the point of intersection of optical axis and the image sensor, two parameters), the center of projection (three parameters), and the camera orientation (three spatial angles). All these parameters in all cameras are calibrated using a specially designed calibration target. The planar target consists of 63 calibration points, each of them of known position (to within 0.1 mm). Each calibration point has an index, which can be automatically identified from its image. A lookup table is used to assign within-plane coordinates to each index. Calibration point indices are coded using self-correcting binary code (see Vuylsteke and Oosterlinck, 1990, for details). The planar target is moved twice by 3 in. in perpendicular direction. Each camera is then automatically calibrated from the three images of the target.

### 2.2. Validation study

We recruited Philadelphia area actors ($N = 70$) and actresses ($N = 69$) in collaboration with Aaron Posner, artistic director of the Arden Theater. They ranged in age from 10 to 85 years (mean $\pm$ S.D. men: $38.5 \pm 14.8$; women $36.4 \pm 16.3$), with proportional ethnic representation (91 Caucasian, 32 African American, six Asian, ten Hispanic). Expressions obtained were happiness, sadness, anger, fear, disgust (Fig. 6), and neutral, representing the expressions that can be reliably rated cross culturally (Ekman and Rosenberg, 1997). The expressions were obtained in two conditions traditionally used in directing (Morgan, 1984): Posed, where the actors were instructed to express 3° of each emotion using a mechanical approach (the English method); Evoked, actors were coached to re-live appropriate experiences using a standard protocol executed by two professional directors. Autonomic measures were obtained as well as mood ratings and videotapes.

In order to validate the detectibility of the intended emotion, we have presented a subset of the faces to successive samples of healthy raters. Data are available on seven faces (three actors, four actresses) rated by 45 raters and an additional eight faces (four actors, four actresses) rated by 62 raters. Examination of the remaining faces is ongoing, but will take longer to complete because it is not feasible to include more than eight faces in a rating session (each face has 40 stimuli to be rated: five emotions × three intensity levels × two conditions + 5 × 2 neutrals). Already, the preliminary data (Fig. 7) indicate that all expressions are detected correctly at levels far better than chance. Furthermore, the variability in accuracy, with happiness being most easily identified and anger the least, is consistent with effects reported in the literature (e.g. Wallbott, 1998). Finally, the effect of evoked relative to posed emotions was evident, with better accuracy for evoked expressions in all emotions except disgust.

To examine which emotions are most often mistaken for which other emotions, we have generated a ‘confusion matrix’ counting the frequency distribution of incorrect responses for each emotion. As can be seen in Fig. 8, the most frequent error for all emotions except disgust was to label them as ‘neutral’ while disgust and neutral were most frequently mislabeled as ‘sad.’ It is also noteworthy that very few faces were mislabeled as happy or fearful.

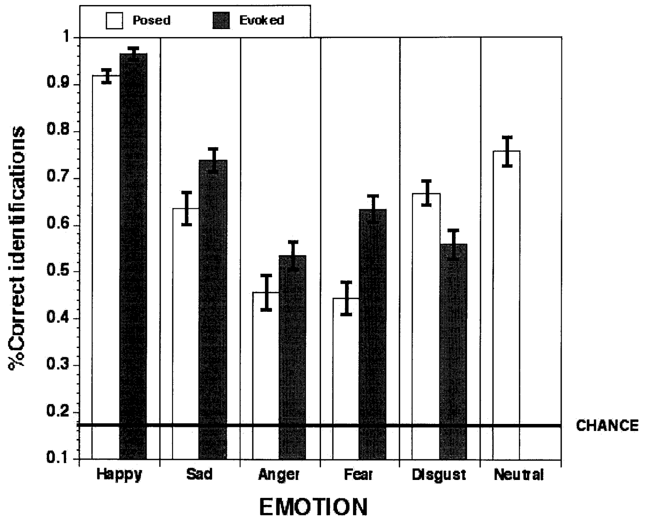

Fig. 7. Percent of correct identification for the five emotions in both posed and evoked conditions and for the neutral expression.

## 3. Discussion

The method described for acquiring 3-dimensional photographs of emotional displays is feasible, produces uniformly good results across the range of age and ethnicity, and shows evidence of validity. The reconstructions are of high quality and can be easily classified by raters with regard to the identity of the emotion displayed. Furthermore, the accuracy of detecting the different emotions is similar to that reported in earlier studies (Erwin et al., 1992; Wallbott, 1998). Finally, a ‘confusion matrix’ of errors did not suggest any systematic misrepresentation of specific emotions. The method can thus be used both for acquiring new facial expressions and for direct application in neurocognitive studies.

The new set of emotion stimuli uses digital image acquisition and processing technology. The primary advantage of the new methodology is the ability to apply image analysis algorithms to construct precise Euclidean models of the face in 3D. This enables quantification of both within-subject and between-subject variability. It also eliminates effects of tilt, which is crucial for examining asymmetry in facial expressions. Other advantages over analog image acquisition include: (1) Data are in frames that can be post-processed and presented individually, as averages, or as a ‘movie’; (2) They are stored in a relational database for rapid access when needed to prepare tasks; (3) They produce a more ‘real’ effect and are likely to be more powerful for mood induction and eliciting physiologic response; (4) Images can be rotated and information enhanced or degraded to control identification difficulty.

The images are available for use by the scientific community. To obtain a set for your use please log on to our website at: http://www.uphs.upenn.edu/bbl/downloads/requests.shtml where you will find sample images and a request form. In the request form, you can provide your research credentials and specify your needs, including the format of images (VRML, MPEG or JPEG). After submitting the request, usually within 48 h (2 working-days) you will receive instructions on how to obtain the images.

The method in its current implementation has several weaknesses. With four digital cameras it is impossible to cover the area under the chin and some parts posterior to the ears. An arrangement with six cameras is more optimal. Other shortcomings include the time it takes to reconstruct the 3-dimensional model, currently it is about 4.5 min on a Pentium III (750 MHz) per image, and the need to acquire an analog picture for color

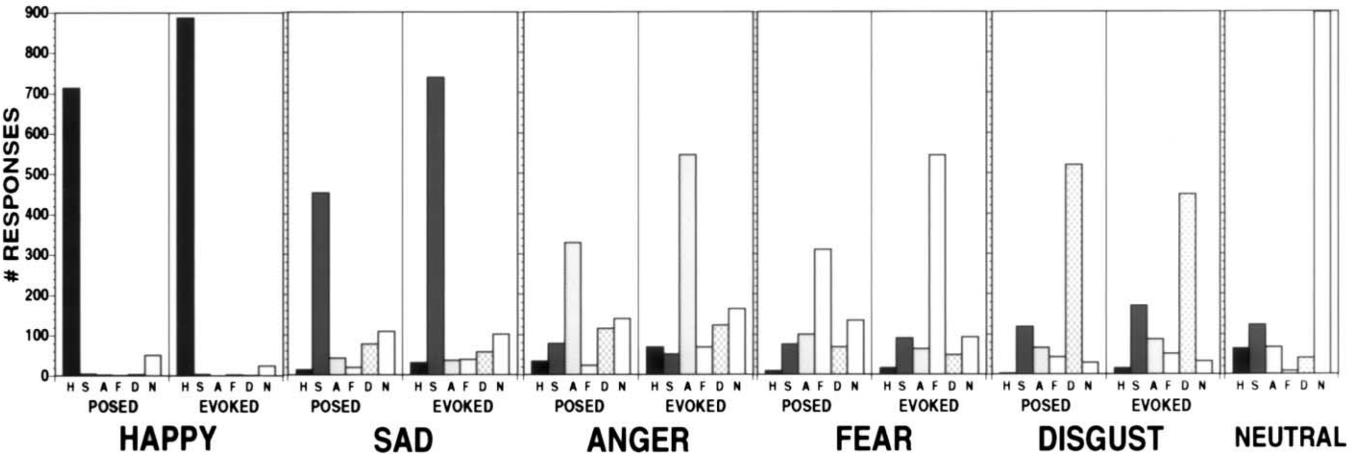

Fig. 8. A ‘confusion matrix’ of the distribution of errors for each of the five emotions in both posed and evoked conditions and for the neutral expression. Errors are H, happy; S, sad; A, anger; F, fear; D, disgust; N, neutral.

rendering. However, these limitations can be overcome with advances in technology.

## Acknowledgements

We thank David Ingraham and Mary Jackman, Directors and the dedicated actors who produced the emotional expressions. We also thank Brian Klimas for assistance. Supported by the Bosworth Fund, NIMH Grants MH60722, MH43880, Grant Agency of the Czech Republic (Project GACR 102/01/1371) and Czech Ministry of Education (Project MSM 212300013).

## References

Adolphs R, Tranel D, Damasio H, Damasio A. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. Nature 1994;372:669–72.

Adolphs R, Damasio H, Tranel D, Damasio AR. Cortical systems for the recognition of emotion in facial expression. J Neurosci 1996;16:7678–87.

Bajcsy R, Reyes E, Kamberova G, Sara R. A 3D Geometric Model Acquisition System for a Tele-Collaboration Testbed. In: Proceedings of ICASE/LaRC/ARO/NSF Workshop on Computational Aerosciences in the 21st Century. Hampton, VA: Kluwer Publisher, 1998.

Blonder LX, Bowers D, Heilman KH. The role of the right hemisphere in emotional communication. Brain 1991;115:1115–27.

Borod J. Brain mechanisms underlying facial, prosodic, and lexical emotional expression: Behavioral evidence from brain-damaged and normal adults. Neuropsychology 1993;7:745–57.

Borod J, Haywood JC, Santschi C, Koff E. Neuropsychological aspects of facial asymmetry during emotional expression: a review of the normal adult literature. Neuropsychol Rev 1997;7:41–60.

Breiter HC, Etcoff NL, Whalen PJ, Kennedy WA, Rauch SL, Buckner RL, Strauss MM, Hyman SE, Rosen BR. Response and habituation of the human amygdala during visual processing of facial expression. Neuron 1996;17:875–87.

Ekman P, Rosenberg EL. What the Face Reveals. NY: Oxford University Press, 1997.

Erwin RJ, Gur RC, Gur RE, Skolnick BE, Mawhinney-Hee M, Smailis J. Facial emotion discrimination: I. Task construction and behavioral findings in normals. Psychiatry Res 1992;42:231–40.

Hyman SE. A new image of fear and emotion. Nature 1998;393:417–8.

Kohler CG, Bilker W, Hagendoorn M, Gur RE, Gur RC. Emotion recognition deficit in schizophrenia: association with symptomatology and cognition. Biol Psychiatry 2000;48:127–36.

McKendall R, Sara R, Bajcsy R. Scalable parallel computing for real time 3D reconstruction from polynocular stereo. http://www.cis.upenn.edu/-mcken/progressll/main.html. 1997.

Morgan JV. Stanislavski’s encounter with Shakespeare’s: The evolution of a method. AnnArbor, MI: UMI Research Press, 1984.

Morris JS, Frith CD, Perrett DI, Rowland D, Young AW, Calder AJ, Dolan RJ. A differential neural response in the human amygdala to fearful and happy facial expressions. Nature 1996;383:812–5.

Morris JS, Friston KJ, Buchel C, Frith CD, Young AW, Calder AJ, Dolan RJ. A neuromodulatory role for the human amygdala in processing emotional facial expressions. Brain 1998;121:47–57.

Phillips M. Geomview Manual. Minneapolis, MN: The Geometry Center, 1993.

Sackeim HA, Gur RC, Saucy MC. Emotions are expressed more intensely on the left side of the face. Science 1978;202:434–6.

Sara R. Accurate natural surface reconstruction from polynocular stereo. In: Leonardis A, Solina F, Bajcsy R, editors. Proceedings NATO Advanced Research Workshop Confluence of Computer Vision and Computer Graphics. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2000:69–86.

Sara R. Sigma-Delta Stable Matching for Computational Stereopsis. Research Report CTU-CMP-2001-25 of Center for Machine Perception, Faculty of EE, Czech Technical University, September 2001.

Sara R, Bajcsy R. Fish-Scales: Representing Fuzzy Manifolds. In: Chandran S, Desai U, editors. Proceedings of the 6th International Conference on Computer Vision. New Delhi, India: Narosa Publishing House, 1998.

Vuylsteke P, Oosterlinck A. Range image acquisition with a single binary-encoded light pattern. IEEE Trans PAMI 1990;12:148–64.

VRML Standards and specifications. http://www.web3d.org/Specifications/.2001

Wallbott H. Big girls don’t frown, big boys don’t cry—Gender differences of professional actors in communicating emotion via facial expression. J Nonverbal Behav 1998;12:98–106.