

A Methodology to Establish Ground Truth for Computer Vision Algorithms to Estimate Haptic Features from Visual Images

Troy L. McDaniel¹, Kanav Kahol¹, Priyamvada Tripathi¹, David P. Smith¹, Laura Bratton¹,
Ravi Atreya², Sethuraman Panchanathan¹

¹Center for Cognitive Ubiquitous Computing, Department of Computer Science and Engineering,
Ira A. Fulton School of Engineering, Arizona State University, Tempe, Arizona, 85287

²University of Arizona, Tucson, Arizona, 85721

Abstract – Humans have an uncanny ability to estimate haptic features of an object such as haptic shape, size, texture and material by visual inspection. A significant computer vision problem is that of estimating haptic features from visual images. While explorations have been made in estimation of visual features such as visual texture, work on estimation of haptic features from video is still in its infancy. We present a methodology to establish ground truth for estimation of haptic features from visual images. We assembled a visio-haptic database of 48 objects ranging from nonsense objects to everyday objects. The variation was controlled in objects by systematically varying haptic features such as shape and texture, and the physical and perceptual ground truth of visual and haptic features was documented. This database provides visio-haptic features of objects and can be used to develop algorithms to estimate haptic features from visual images. Finally, a tactile cueing experiment is presented demonstrating how visio-haptic ground truth can be used to assess the accuracy of a system for visio-haptic conversion of image content.

Keywords – Image databases, Testing, Tactile displays, Tactile systems, Machine vision, Visualization

I. INTRODUCTION

Haptics refers to the science of touch. Touch as a modality is a basic sensory mechanism in humans specialized to perceive spatial properties of objects and environments [5]. Incidentally, the human visual system is also geared towards spatial perception and there is a considerable overlap between the visual and haptic modality. This overlap of information gives humans the ability to estimate haptic features of an object by its image. The process of estimating what an object feels like from its visual image is a probabilistic process that involves neural circuitry for intermodal coordination in the brain employing rules to estimate features in one modality based on input from another.

Computer vision algorithms for visio-haptic conversion attempt to extract haptic data directly from images or extract visual data that is subsequently converted into haptic data using intermodal coordination rules. Visio-haptic research spearheads efforts to provide haptic augmented reality environments where virtual haptic objects are recreated from visual information extracted from objects in real environments. Such a technology has many important applications including the development of assistive devices for perceiving distal environments through haptic user

interfaces for individuals who are blind. Other applications include teleoperation and telesurgery.

Some effort has been made towards automatic conversion of visual data into haptic features [2, 6-9]. It may be noted that many of these efforts have focused on the extraction of visual features such as shape and texture, and assume the validity of these features in the haptic modality. However, visio-haptic conversion is a probabilistic process that in certain cases may not lead to perceptually accurate results. For example, visio-haptic contradictions exist in a sponge colored to look like a rock. This suggests that visual and haptic features are perceived and processed disjointedly in the human brain. Hence, a simplistic approach involving the assumption that haptic features are equivalent to visual features is not adequate to model intermodal visio-haptic coordination rules.

Existing methods for estimation of haptic features from visual images have limited applicability due to a lack of ground truth. In the real world, ground truth in any modality can be of two types. Physical ground truth involves measurable quantities for features such as volume for shape and roughness grade for texture. However, the world is perceived in terms of perceptual concepts rather than physical data. Perceptual ground truth involves subjective evaluations of features such as shape, size, texture and material, and classification based on these subjective measures. Object databases need to provide both physical and perceptual ground truth to facilitate development of algorithms that link physical and perceptual ground truth. However, existing databases for testing computer vision algorithms are mainly designed for visual object recognition and visual texture classification tasks. Available databases do not include both perceptual and physical ground truth for objects, and document the visual modality only. There is no available database for haptic or visio-haptic properties severely limiting visio-haptic research.

In this paper, we present a methodology to assemble a comprehensive database of objects that enlists the objects' physical and perceptual ground truth in both visual and haptic modalities. This database can be utilized for systematic explorations in developing computer vision algorithms and human-computer interfaces that estimate haptic features such as shape, texture, material and size from visual data. Results from a tactile cueing experiment are documented

demonstrating how visio-haptic ground truth can be used to assess the accuracy of algorithms for visio-haptic conversion.

II. BACKGROUND AND RELATED WORK

Research teams have developed databases of objects, scenes and textures providing a framework for testing the accuracy of computer vision algorithms. We conducted a survey on popular image processing and computer vision databases to investigate current limitations. Databases were inspected based on the following criteria: *image count, image format, registered or unregistered, color or grayscale, single or multiple objects, synthetic or real, quantified measurements, variations in features employed, availability of stereo images, availability of visual ground truth, and database accessibility*. None of the currently available databases provide haptic ground truth, and only some provide visual ground truth. The survey can be downloaded from [1].

The work done on automatic visio-haptic conversion of image content has primarily focused on 2D-to-3D conversion or the tactization of edges or illumination values. A popular system is the tactile vision sensory substitution system (TVSS) [2]. TVSS is a wearable system that converts illumination values into vibrotactile stimulations delivered using a 20-by-20 tactile pin array. This system assumes a direct mapping from the visual modality to the haptic modality and presents haptic information at a sensory rather than perceptual level, creating a system that is difficult to use, and thus requiring extensive training. For example, an average of 10 hours of training was required to recognize a single basic object at different orientations.

In [9], researchers developed TACTICS, a visual-to-tactile image translator designed for offline interpretation of images. This system extracts edges from an image and subsequently creates a raised line image. Again, this system assumes a direct mapping from the visual modality to the haptic modality, resulting in a haptic presentation of visual information that is not very effective. Some usability tests are performed with participants, however the most appropriate method for testing the effectiveness of an algorithm for visio-haptic conversion is an evaluation based on ground truth. Little or no algorithmic verification against ground truth pervades the related literature including [6-8] due to a lack of available visio-haptic ground truth, limiting progress of visio-haptic conversion of image content.

III. CONCEPTUAL FRAMEWORK

Capturing physical and perceptual ground truth of visual and haptic features of objects requires a consistent methodology to acquire and document the properties of objects. Table 1 depicts the framework for ground truth collection. Capture is divided into physical and perceptual capture, and these are further divided into shape, size, texture and material capture.

For shape and size ground truth in the visual modality, we captured stereo images of the objects. Visual texture and

material are captured through images of the texture and material in close proximity. Visual inspection and classification by individuals is used to generate the perceptual ground truth. This involves participants classifying object features into predetermined classes.

Collecting physical ground truth of haptic data poses a special challenge since tactile sensor technology is in its infancy. Currently our database does not include physical ground truth for haptic texture and material since texture and material sensors that can document features of everyday objects are not currently available. Haptic shape and size are captured by moving a 3D tracker over an object’s surface and recording $\langle x,y,z \rangle$ coordinates at 30 frames/sec.

Table 1. Framework for ground truth collection

Physical Ground Truth (Visual Modality)	
Shape, Size	Stereo Imaging
Texture, Material	Visual Capture
Perceptual Ground Truth (Visual Modality)	
Shape, Size, Texture, Material	Visual inspection and classification
Physical Ground Truth (Haptic Modality)	
Shape, Size	3D tracking of surface
Perceptual Ground Truth (Haptic Modality)	
Shape, Size, Texture, Material	Haptic exploration and classification

Research in psychology of haptic perception suggests that perception and action are closely related in the haptic modality [5]. Various attempts have been made to study the manual exploratory procedures of individuals who are blind or sighted. A seminal study of exploratory procedures was reported by Lederman and Klatzky [5]. They asked adults to use haptic exploration to classify objects, according to a given criterion. This allowed them to identify specific exploratory hand movements (which they called “exploratory procedures”) which were characterized by (1) the quantity and the nature of the information that each procedure provided, and (2) the range of properties for which each procedure was useful. Lederman and Klatzky reported that many of the exploratory procedures used by their participants were related to the object property being explored, in a one-to-one relation. For example, certain subjects always used lateral motion for perceiving texture.

Lederman and Klatzky also noted that there are two distinct phases of the object exploration procedures in human adults. During the first phase, they employ generalized procedures that mobilize the whole hand, and gather vague haptic and tactile information about several properties. During the second phase, specific exploratory procedures are used to perceive particular object features. This tends to suggest that haptic exploration of objects quantified as hand gestures is a useful measure of perceived features and their quantification values.

In the design of our perceptual ground truth collection for the haptic modality, we asked participants to classify object features while we capture their 3D hand movements. Haptic explorations given by the 3D hand movement profiles are indicative of haptic features being explored and the

perceptual salience of features. This is similar to eye saccade movement used in visual perception. Haptic exploration, however, is more informative as the type of movement indicates the feature being perceived.

The framework for algorithmic evaluation is straightforward. Visual ground truth can verify the accuracy of algorithms for visual information extraction before visio-haptic conversion. Once conversion has completed, haptic ground truth can verify the accuracy of conversion. Scores can then be combined using a weighted hierarchical accuracy measure (Eqn. 1-3).

$$\text{Final score} = a_1PH + a_2PE \quad (1)$$

where

$$PH = b_1PH_v + b_2PH_h, \quad PE = c_1PE_v + c_2PE_h \quad (2)$$

where

$$\begin{aligned} PH_v &= d_1PH_{v,sh} + d_2PH_{v,sz} + d_3PH_{v,tx} + d_4PH_{v,ma} \\ PH_h &= d_1PH_{h,sh} + d_2PH_{h,sz} + d_3PH_{h,tx} + d_4PH_{h,ma} \\ PE_v &= d_1PE_{v,sh} + d_2PE_{v,sz} + d_3PE_{v,tx} + d_4PE_{v,ma} \\ PE_h &= d_1PE_{h,sh} + d_2PE_{h,sz} + d_3PE_{h,tx} + d_4PE_{h,ma} \end{aligned} \quad (3)$$

In Eqn. 1-3, PH and PE denote physical and perceptual scores. In Eqn. 3, subscripts v and h denote visual and haptic modalities, and subscripts sh , sz , tx and ma denote shape, size, texture and material. All scores must be between 0 and 1, inclusive. Weights a_i , b_i , c_i and d_i are chosen by the user, where $a_1+a_2=1$, $b_1+b_2=1$, $c_1+c_2=1$, and $d_1+d_2+d_3+d_4=1$. The first level, Eqn. 1, combines the physical and perceptual scores. The second level, Eqn. 2, combines scores in the visual and haptic modalities. The third level, Eqn. 3, combines shape, size, texture and material scores.

IV. EXPERIMENTAL METHODOLOGY

Our visio-haptic object database, depicted in Fig. 1, consists of forty-eight objects. Twelve of the objects are irregular objects, made from legos. The other thirty-six objects are everyday objects, consisting of twelve bowls, twelve cups and twelve glasses. We controlled variations in haptic shape, size, texture and material.



Figure 1. Visio-haptic object database

Our experimental methodology is divided between the capture of visual and haptic features of objects. We also conducted a tactile cueing experiment demonstrating how visio-haptic ground truth can be used to evaluate algorithms for visio-haptic conversion. The complete database including

test images, physical and perceptual ground truth in the visual and haptic modality, and measurements for each capture setup, can be downloaded from [1].

A. Visual Capture

The capture setup for test image acquisition is shown in Fig. 2a. Objects were placed on top of an elevated platform, 1.23 meters high, resting on level ground, which is located in the center of a dual rotating platform, 0.025 meters thick and 8 meters in diameter. A Canon NTSC 2R20 Digital Video Recorder on a tripod is mounted on the rotating platform. The tripod together with the camera is 1.23 meters high, and is 0.65 meters in front of the center of the platform. Opposite from the camera on the rotating platform is a white backdrop, 1.5 meters in front of the camera. The zoom of the camera is such that two marks on the white backdrop are a distance of 0.75 meters apart when they appear in the upper corners of the video. For the tall glasses (nine total), a larger viewing area is required, and thus marks have a distance of 0.90 meters.

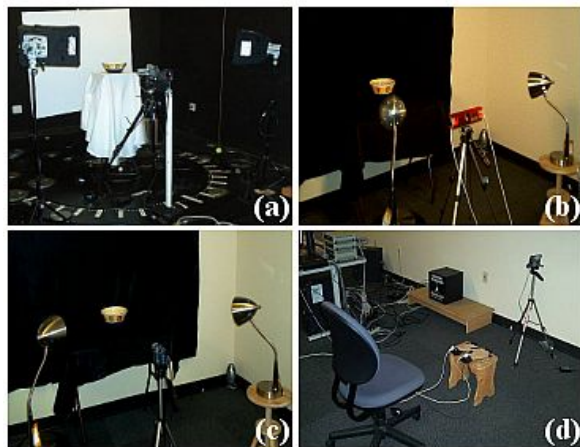


Figure 2. Capture setups: (a) test images, (b) stereo, (c) material/texture, (d) perceptual (haptic modality)

Two Photoflex Starlite diffused light sources are located in the foreground of the capture setup. From the viewpoint of the object, the right light is at a height of 1.5 meters, has a diffuse light intensity of EV 12, and is at an angle of -45° . Similarly, the left light is at a height of 1.46 meters, has a diffuse light intensity of EV 13, and is at an angle of 45° . These are the only light sources used during capture. The ambient light intensity at the center of the platform is EV 7.

The camera starts at -90° and rotates to $+90^\circ$, recording video at 30 frames/sec. The video is streamed to the workstation for storage. We performed this process for each object, capturing 48 videos.

Frames are extracted from the videos at an interval of one degree. The frames are cropped and padded to create a database of same-sized images. Sobel edge detection is utilized to find the bounding box of objects for segmentation;

linear interpolation is used to pad the images to a registered size.

For visual ground truth, we gathered physical shape, size, texture and material in the visual modality. Shape and size information was acquired using variable-baseline stereoscopic equipment and the Small Visions System (SVS) software purchased from Videre Design. We captured 360 degrees at 5-degree intervals of each object by rotating objects on a platform in front of the stereo rig. Data is collected in the form of stereo pairs and shape information consisting of 3D coordinates for each object view. Fig. 2b shows the capture setup. For setup measurements see [1].

Texture and material ground truth was collected through visual capture of texture and material by taking images at close range of objects under various lighting conditions. The capture setup is shown in Fig. 2c. For setup measurements see [1].

Perceptual capture of visual information was collected through an experiment involving twenty sighted participants, mostly students from our lab with no vision problems that would be detrimental to the results. Each participant sat at a table, while twelve different objects selected from our database were placed in front of him or her. Three objects are selected from each category: irregulars, bowls, cups, and glasses. The order of categorical selection is random. A row (1, 2, 3 or 4) within each category is selected at random, but such that row numbers do not repeat for each participant and each particular row has not yet been explored within the current set of four participants. This allows each object to be explored five times over the course of the entire experiment. The majority answer was taken as the consensus for each question.

Participants were asked to visually inspect each object without touching it to answer a series of questions. An object is divided into six regions: the rim, base, inside, lower vertical region (LV), middle vertical region (MV), and the upper vertical region (UV). Participants answered a total of twenty-eight questions, four questions for the overall object and each object region except the inside region which has only two questions. Furthermore two questions were asked pertaining to auxiliary information.

The multiple choice questions for each region or overall object pertain to that particular region’s shape, size, texture and material. The questions are presented in Table 2. The same questions were presented during perceptual capture of haptic ground truth as well.

B. Haptic Capture

Physical shape and size information in the haptic modality was acquired using Ascension® 3D trackers. Tracking equipment allows continuous extraction of the trackers’ $\langle x,y,z \rangle$ coordinates as they move in space. Moving the tracker along the surface of an object creates a 3D point cloud representing the object’s shape and volume. Researchers can subsequently apply algorithms to these point clouds to

determine the parameters of a fitted surface for shape, or analyze max-min points for size.

Perceptual haptic ground truth was collected through an experiment involving five blind and fifteen sighted (blind-folded) participants. Most of the sighted participants were students from our lab, and most of the blind participants were students at ASU. Participants that had already participated in the visual experiment were given different objects than the ones before.

Participants wore CyberTouch® gloves by Immersion®, which were used in conjunction with Ascension® 3D trackers allowing hand movements to be recorded as objects are haptically explored. The system records finger joint angles, thumb and pinkie rotations, wrist pitch and yaw, and the $\langle x,y,z \rangle$ coordinates of finger joints and each wrist. The capture setup is shown in Fig. 2d.

Table 2. Perceptual ground truth questionnaire

Question	Answers
Overall Shape	a) Bowl b) Cup c) Glass d) Irregular
Overall Size	a) Small b) Medium c) Large
Overall Texture	a) Smooth b) Medium c) Rough
Overall Material	a) Plastic b) Metal c) Ceramic d) Glass e) Cloth f) Other:
Base Shape	a) Round b) Rectangular c) Irregular
Base Size	a) Small b) Medium c) Large
Base Texture	a) Smooth b) Medium c) Rough
Base Material	a) Plastic b) Metal c) Ceramic d) Glass e) Cloth f) Other:
Rim Shape	a) Round b) Rectangular c) Irregular
Rim Size	a) Small b) Medium c) Large
Rim Texture	a) Smooth b) Medium c) Rough
Rim Material	a) Plastic b) Metal c) Ceramic d) Glass e) Cloth f) Other:
Inside Shape	a) Round b) Rectangular c) Irregular
Inside Texture	a) Smooth b) Medium c) Rough
LV Curvature	a) None b) Low c) Medium d) High
LV Size (diameter)	a) Small b) Medium c) Large
LV Texture	a) Smooth b) Medium c) Rough
LV Material	a) Plastic b) Metal c) Ceramic d) Glass e) Cloth f) Other:
MV Curvature	a) None b) Low c) Medium d) High
MV Size (diameter)	a) Small b) Medium c) Large
MV Texture	a) Smooth b) Medium c) Rough
MV Material	a) Plastic b) Metal c) Ceramic d) Glass e) Cloth f) Other:
UV Curvature	a) None b) Low c) Medium d) High
UV Size (diameter)	a) Small b) Medium c) Large
UV Texture	a) Smooth b) Medium c) Rough
UV Material	a) Plastic b) Metal c) Ceramic d) Glass e) Cloth f) Other:
Taller Than Wider	Yes/No
Base Larger Than Rim	Yes/No

Participants were handed twelve different objects to haptically explore. Objects were selected from our database using the same selection process used during visual capture. The questions listed in Table 2 were asked to each participant. Similar to visual capture, the majority answer was taken for each question. During haptic exploration, hand movements were recorded both from the gloves and from a video camera for additional offline analysis. Participants were

instructed to make a fist gesture between questions to facilitate automatic segmentation of hand movements. Hand movement data are included in the database as ground truth.

C. Tactile Cueing Experiment

To test the usefulness of our database and demonstrate how it can be applied to assess the accuracy of algorithms for haptic feature estimation from visual images, we developed an application that converts images of bowls, cups and glasses to tactile cues that convey information about shape, size and texture. Presentation of haptic features as vibrotactile cues offers a promising alternative to realistic haptic rendering in that tactile cues can quickly invoke a mental image of an object in a user’s mind. Furthermore, hardware for cueing is a lot less expensive compared to force-feedback devices for realistic haptic rendering.

An object’s shape is classified using principal component analysis (PCA) in terms of ‘bowl’, ‘cup’ or ‘glass’ quantization values. Size classification is performed by fitting a bounding box around an object, then perceptually categorizing its area (at a fixed distance) with respect to its identified object category. Quantization values for size are small, medium and large. Finally, texture classification is performed by perceptually categorizing the standard deviation of a sub-window containing an object’s texture. Quantization values for texture are smooth, medium and rough. Perceptual categorizations of shape, size and texture are then conveyed to a user through tactile cues using CyberTouch® gloves by Immersion®.

Previously, we worked with a participant who is blind to develop a tactile cueing language [4]. The participant assigned cues to each of the perceptual values for shape, size, texture and material, as well as auxiliary information including questions such as “Is the object taller than it is wider?” and “Is the base larger than the rim?” The participant was trained on five objects from our visio-haptic object database. With only approximately 0.5 hours of training, the participant was able to recognize every object in our database (48 objects total) given the correct cues with an impressive recognition accuracy of 100% [4]. These results suggest that tactile cues can be learned quickly and retained, and are very effective at invoking mental concepts within a user’s mind.

V. RESULTS

All forty-eight objects have been captured and their ground truth collected (available for download from [1]) following the methodology outlined in this paper. Compared to other databases used for testing computer vision algorithms, our database provides (a) physical and perceptual ground truth in both the visual and haptic modalities, (b) same-size (registered) images and (c) quantification of the setups, allowing repeatability. The physical ground truth for one of the forty-eight objects is depicted in Fig. 3, along with its test image. The perceptual ground truth in the haptic modality for this object is given in Table 3. (The visual

modality differs only by the size of the lower vertical region, which is small.)

We trained our system on a subset of the visio-haptic object database, and tested using the remaining database objects (sixty images from six novel objects). The participant who trained on our database decided that the most salient features were shape, size and texture, in that order, agreeing on weights of 0.5, 0.3 and 0.2, respectively. We gave more weight to the haptic modality, i.e., 0.8 (haptic) and 0.2 (visual), and a weight of 1.0 to the perceptual score since our system avoids extracting physical information, and instead, classifies simple features into perceptual values. Also, a direct mapping between the visual and haptic modality is assumed. Using scores calculated for shape, size and texture classification (discussed next) and user-specific weights, Eqn. 1-3 gives our system an overall accuracy of 88%.

Comparing results for shape information against perceptual ground truth, we achieved 100% accuracy for both modalities. Comparing results for size information against perceptual ground truth, we achieved 83% (haptic) and 68% (visual) accuracy. As expected, results for texture classification yielded poor results due to the use of a visual texture rather than haptic texture algorithm: 68% (haptic) and 68% (visual). Instead, algorithms for haptic texture analysis, such as texture-from-shading, need to be developed and utilized.

VI. CONCLUSION AND FUTURE WORK

This paper presents a methodology for acquiring and documenting visual and haptic features based on physical and perceptual spatial properties. This methodology was used to develop a novel object database that can be used to assess the accuracy of visio-haptic conversion of image content. We are currently in the process of developing the tactile sensor technology to aid in texture and material physical ground truth collection. The database is also being augmented with similarity measures in the visual and haptic modality of the various object pairs. This data can be employed to explore the visual and haptic perceptual space and study their relation.

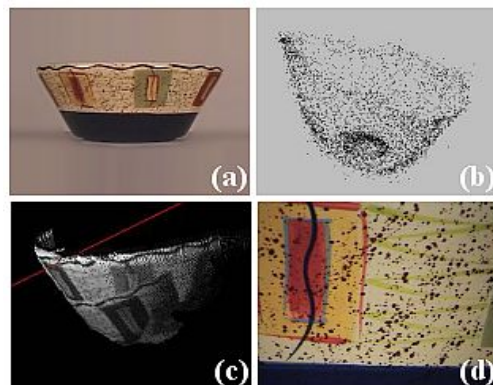


Figure 3. Test image and physical ground truth: (a) test image, (b) haptic shape/size, (c) visual shape/size, (d) visual texture/material

Table 3. Perceptual ground truth (haptic modality)

Overall	Bowl	Large	Smooth	Ceramic
Rim	Round	Medium	Smooth	Ceramic
Base	Round	Large	Smooth	Ceramic
Inside	Round	N/A	Smooth	N/A
LV	Medium	Medium	Smooth	Ceramic
MV	Medium	Medium	Smooth	Ceramic
UV	Medium	Large	Smooth	Ceramic

VII. REFERENCES

- [1] <http://cubic.asu.edu> (data under resources link)
- [2] P. Bach-y-Rita, *Brain mechanisms in sensory substitution*. Academic Press, New York, 1972.
- [3] R.S. Johansson and A.B. Vallbo, "Tactile sensory coding in the glabrous skin of the human hand", *Trends in Neuroscience*, 6(1), 1983, pp. 27-32.
- [4] K. Kahol, P. Tripathi, and S. Panchanathan, "Haptic User Interfaces: Design, testing and evaluation of haptic cueing systems to convey shape, weight, material and texture information", presented at *International Conference on Human-Computer Interfaces*, Las Vegas, Nevada, May 2005.
- [5] S.J. Lederman and R.L. Klatzky, "Hand movements: A window into haptic object recognition", *Cognitive Psychology*, Vol. 19, 1987, pp. 342-368.
- [6] P. Roth, D. Richoz, L. Petrucci, and T. Pun, "An audio-haptic tool for non-visual image representation", *Proc. of ISSPA '01*, Kuala Lumpur, Malaysia, 13-16 August 2001, pp. 64-67.
- [7] Y. Shi and D.K. Pai, "Haptic display of visual images", *Proc. of VRAIS '97*, Albuquerque, New Mexico, Mar. 1997, pp. 188-191.
- [8] B. la Torre, D. Prattichizzo, F. Barbagli, and A. Vicino, "The FeTouch project", *Proc. of IEEE ICRA '03*, Taipei, Taiwan, 14-19 Sept. 2003.
- [9] T.P. Way and K.E. Barner, "Automatic Visual to Tactile Translation, Part II: Evaluation of the TACTile Image Creation System", *IEEE Transactions on Rehabilitation Engineering*, March 1997, pp. 95-105.