# Bayesian Tactile Face

Zheshen Wang, Xinyu Xu, Baoxin Li
Computer Science and Engineering
Arizona State University, Tempe, AZ, 85281, USA
{zheshen.wang, xinyu.xu, baoxin.li}@asu.edu

## Abstract

*Computer users with visual impairment cannot access the rich graphical contents in print or digital media unless relying on visual-to-tactile conversion, which is done primarily by human specialists. Automated approaches to this conversion are an emerging research field, in which currently only simple graphics such as diagrams are handled. This paper proposes a systematic method for automatically converting a human portrait image into its tactile form. We model the face based on deformable Active Shape Model (ASM)[4], which is enriched by local appearance models in terms of gradient profiles along the shape. The generic face model including the appearance components is learnt from a set of training face images. Given a new portrait image, the prior model is updated through Bayesian inference. To facilitate the incorporation of a pose-dependent appearance model, we propose a statistical sampling scheme for the inference task. Furthermore, to compensate for the simplicity of the face model, edge segments of a given image are used to enrich the basic face model in generating the final tactile printout. Experiments are designed to evaluate the performance of the proposed method.*

## 1. Introduction

With the text-to-speech/text-to-Braille technology, nowadays computer users with visual impairment can independently access digital textual information. However, when it comes to digital graphical contents, there still exist major barriers for a computer user who is blind to independently access the vast amount of digital graphical media such as graphs, diagrams and images. Traditional approaches to solving this problem rely on third-party sighted professionals, tactile graphic specialists (TGS), to manually convert the digital graphics into their tactile forms. This process is typically time-consuming and labor-intensive. Thus it is imperative to develop automatic approaches that can convert visual digital graphics into their tactile forms so that the computer users with visual impairment may gain immediate access to graphical contents with such technologies.

Due to the extremely limited bandwidth of tactile perception compared with that of vision, image simplification is a key step in automatic visual-to-tactile conversion. Unfortunately, in the literature, this key step has not been fully studied. The work of [1] relied on simple image processing steps such as negation and edge detection. Way et al [2] proposed to simplify images also mostly by edge detection. However, without high-level guidance, it is difficult to choose the best parameters for such low-level image processing steps. For example, major information, which should be kept, may be lost due to the failure of edge detection. There are also additional problems of broken edges or scattered edge segments that may serve only to confuse a blind user if they are simply mapped to tactile lines as done in [1, 2]. Any attempt to clean up the edges, such as linking short ones to form a long contour, may do harm if those processing steps are purely driven by the data. The system developed in [3] resorts to Photoshop for image simplification, which actually involves the manual efforts of a sighted person and therefore does not address the need of automated solutions that support independent access of a blind user.

To overcome these drawbacks, in this paper, we propose a systematic approach to automatic conversion of a human portrait image into its tactile form. By limiting ourselves to the specific yet commonly-encountered problem, we are able to exploit various constraints imposed by this high-level knowledge of knowing the image containing a face. In particular, instead of simplifying the results from edge detection, we explicitly model human faces using deformable Active Shape Model (ASM) [4]. Prior shape and appearance statistics of human faces are first learned from a set of training face images. Given an image with human portrait, the set of deformable parameters that best explain the image data are estimated through Bayesian inference with statistical sampling. This leads to a Bayesian framework to the tactile conversion task. Upon the match of the face model to the image, edge segments are selectively used to enrich the representation of the aligned face shape, which is by design very simple. Furthermore, in generating the tactile graphics, we exploit the strength of the gradient to modulate the tactile patterns to add desired layer of complexity to the tactile lines. As such, the proposed method combines the model generality with data

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

specificity to create an informative tactile representation of the original face image. To the best of our knowledge, this is the first reported systematic attempt to this challenging problem of automatically creating tactile faces from images.

In Section 2, we briefly discuss related work. In Section 3, we describe the shape and appearance model and also analyze the advantage of the proposed appearance model based on clustering of the gradient profiles along the shape model. Section 4 presents our method for updating a prior model with the given data, including the steps for creating the final tactile graphics by enriching the simple face model with detected edge segments. We report systematic evaluation of our method in Section 5, and then conclude in Section 6 with discussion on future work.

## 2. Related work

Visual-to-tactile conversion has been an active research field recently. Currently active work focuses primarily on simple line-drawing graphics only [5,6], with little effort dealing with natural images such as portrait images that we attempt to address in this paper.

Face alignment is an active research area with many research papers in recent years. In the pioneering work ASM [4], each feature point is sampled by a local profile search, and then a Gaussian shape prior model obtained from PCA is used to regularize the search results iteratively. Bayesian Tangent Shape Model (BTSM) [7] is another derivation of ASM proposed to infer shape parameters more accurately and robustly by the EM algorithm. While being useful, ASM has some known drawbacks. First, ASM tends to get stuck in local minima [8][9] due to the fact that it considers the local optimization of shape points independently [9]; Second, it has difficulty with generalization [10] because using PCA often restrict the local deformations too much.

A number of approaches have been proposed to remedy these problems. For the first problem, a hierarchical CONDENSATION is proposed in [11] to search the MAP estimates of shape configurations. Coughlan et al [12] introduced Markov Random Field model into face alignment to model both the local image structure and the shape prior. Liang et al [13] integrated the Markov network search with the global shape prior to achieve more accurate alignment results but the shape prior is still built by PCA. Huang et al [9] generate new shape by maximizing local probability that describes how desirably the component shape parameters fit the observation, given a probability distribution function (PDF) encoding the interrelationship among parameters of all components modeled by Constrained-GPLVM [14]. To address the second problem, Jiao et al. [8] suggest using Gabor wavelet features to represent the local appearance. Hu et al. [15] utilized a wavelet network to replace the PCA-based appearance model, and demonstrate improved alignment under illumination changes and occlusions. In [10], face alignment is treated as a process of maximizing the score of a trained two-class classifier that is able to distinguish correct alignment from incorrect alignment.

Our approach addresses those two problems from three aspects. First, we employ importance sampling to find the shape parameters that maximize the posterior PDF in a Bayesian formulation, attempting to avoid getting stuck at local minima. Second, as opposed to making the assumption that the prior appearance model is multivariate Gaussian, we perform clustering in learning the appearance model from the training set. This effectively results in a multi-modal representation, which turns out to be able to generalize better. The generalization difficulty of ASM is also due to the fact that an instance of shape configuration is compared with only the appearance variations learnt from the training set, without considering pure data evidence of the current image. This will force the shape to only deform in the ways learnt from the training set. Thus as the third remedy, we combine model-driven part with data-driven part in the likelihood function in our approach. The data-driven term measures how well the global face contour fits to the true face contour, and the model-driven term measures how well the facial features conform to the prior knowledge of shape configurations obtained from the training set. This improvement increases the generalization ability of the algorithm to unseen images.

## 3. Prior shape and appearance model

### 3.1. Statistical shape model

Let $\mathrm{E} = \{(x, y) \in \Re^2\}$ be the image plane. Human faces are characterized by $N$ landmark points [4], $p_i=(x_i, y_i)$, $i=1,...N$. In the IMM [16] face image database used in our system, the following facial features were manually annotated using 58 landmarks around the eyebrows, the eyes, the nose, the mouth, and the jaw (Fig. 1). Seven point paths were used in total, three closed and four open. The shape examples are aligned into a common coordinate frame using Generalized Procrustes Analysis (GPA) [17].

Let $\mathbf{f_i}= (x_{i1}, y_{i1}, x_{i2}, y_{i2}, ..., x_{iN}, y_{iN})^T$ be the feature vector that represents the *i-th* face image in the training set. Given $M$ training images, we form matrix $\mathbf{F}=[\mathbf{f_1}, \mathbf{f_2}, ..., \mathbf{f_M}]$. We want to seek a parameterized model of the form $\mathbf{f}=\Psi(\mathbf{b})$ from the training set, where $\mathbf{b}$ is the parameter of the shape model, such that new shapes can be synthesized using this model by varying $\mathbf{b}$ within reasonable limits. Specifically, we apply PCA to the positions of the landmarks over the training set, $\mathbf{F}$, to find the eigenvectors $\varphi_i$ and the eigenvalues $\lambda_i$ of the covariance matrix $\mathbf{S}$ of $\mathbf{F}$, as done in ASM [4,18]. Then a shape in the training set can be approximated using the mean shape and a weighted sum of the first $t$ largest eigenvectors, i.e.

$$\mathbf{f} \approx \bar{\mathbf{f}} + \Phi\mathbf{b} \quad (1)$$

where $\bar{\mathbf{f}}$ is the mean shape. $\mathbf{\Phi}=[\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2,..., \boldsymbol{\varphi}_t]$ is the matrix of the first $t$ eigenvectors, and $\mathbf{b}=(b_1, b_2,..., b_t)^T$ given by $\mathbf{b} = \Phi^T(\mathbf{f} - \bar{\mathbf{f}})$. The vector $\mathbf{b}$ defines a set of parameters of a deformable model. Eq. (1) allows us to generate new examples of the shapes by varying the elements of $\mathbf{b}$ within suitable limits, such as done in [4] by

$$-3\sqrt{\lambda_k} \le b_k \le 3\sqrt{\lambda_k} \quad (2)$$

In order to reduce the dimensionality of the parameter space, we only use the first $t$ eigenvectors and the first $t$ elements in $\mathbf{b}$, $t$ is chosen such that $p$ percent of the variation is retained, i.e. $\sum_{i=1}^{t}\lambda_i \ge \frac{p}{100}\sum\lambda_i$.

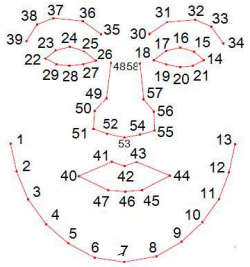


Figure 1: Facial landmarks in the face model [16].

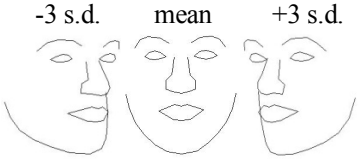
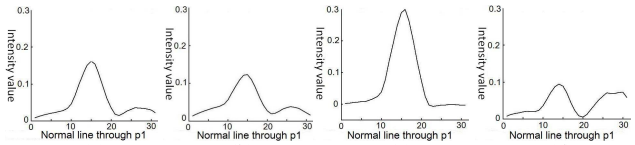



Figure. 2: Effect of varying $b_1$ between ±3 s.d.

Fig. 2 shows the effect of varying the first element of $\mathbf{b}$ in turn between ±3 standard deviations from the mean value, leaving all other elements of $\mathbf{b}$ at zero. The corresponding shape $\mathbf{f}$ is reconstructed using Eq. (1).

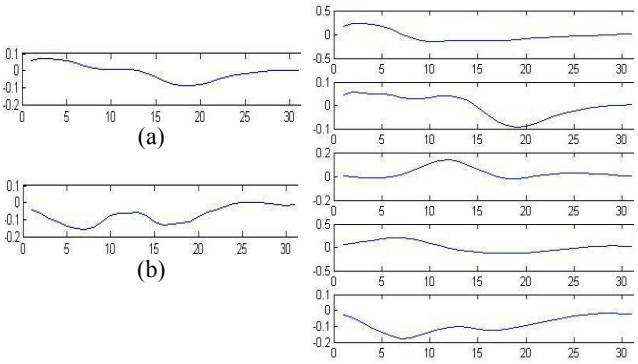
### 3.2. Local appearance model

The local appearance models, which describe local image features around each landmark, are modeled as the first derivative of the profile, $g_j$, computed along the line perpendicular to the boundary of a shape instance $f_i$ through landmark point $p_j$ [4]. Traditional approaches [4][8] assumed the local appearance are distributed as multivariate Gaussian, and the quality of fitting a gradient profile $g_s$ at test image location $s$ to the *j-th* model is computed as the Mahalanobis distance from $g_s$ to the *j-th* model mean. We found a few shortcomings with this method. First, one multivariate Gaussian may not capture the large number of pose and expression variations present in the training set. Second, the model mean is computed across *all the training images*, but the mean gradient profile is very sensitive to the pose (turn left/turn right). For instance, the mean gradient profile of $p_1$ computed over examples that turn left could dramatically differ from the mean gradient profile of $p_1$ computed over examples that turn right, as being verified by Fig. 3. Hence the distinguishable features will be averaged out if the mean gradient profile is computed across different pose set.




(a) (b) (c) (d)

Figure 3: The mean gradient profile is very sensitive to pose. Mean gradient profile at $p_1$ computed over all the training images (a), over the nearly frontal pose set (b), over the turning left pose set (c) and over the turning right pose set (d) look quite different from each other.



(a) (b) (c)

Figure 4: (a) The mean gradient profile at $p_{51}$ computed over all the training images in subset Γ. (b) The gradient profile at $p_{51}$ for a testing image. (c) The 5 cluster centroids obtained from clustering in subset Γ. The score of matching (b) to (a) would be very low, whereas (b) matches well to the 5th cluster centroid in (c), indicating that *k*-means clustering can capture more appearance variations in the training set than simply using the mean gradient profile.

To remedy these shortcomings, we adopt a pose-dependent local appearance model. Specifically, we divide the entire training image set into three subsets based on the pose variation:

$$\begin{aligned} &\{I \in \Gamma \text{ if the face in image } I \text{ is nearly frontal}\} \\ &\{I \in \Theta \text{ if the face in image } I \text{ turns left}\} \\ &\{I \in \Lambda \text{ if the face in image } I \text{ turns right}\} \end{aligned} \quad (3)$$

Further, for each landmark $p_j$, we apply *k*-means clustering to all the gradient profiles of $p_j$ in set Γ, Θ and Λ respectively, keeping 5 clusters for each set. We record the 5 cluster centroids (for each set) and denote them as $\bar{g}_{vj}(\Gamma)$, $\bar{g}_{vj}(\Lambda)$, $\bar{g}_{vj}(\Theta)$, for $v$=1,…, 5, and $j$=1,…, $N$. Fig. 4 illustrates that *k*-means clustering can capture more appearance variations in the training set than simply using the mean gradient profile. For a hypothesized face shape, we first compute which pose set the image belongs to based on the value of $b_1$, then the fitting score at testing landmark $p_j$ is computed as the exponential of the minimum distance between $g_j$ and the 5 centroids,

$$h: b_1 \rightarrow \textit{pose set}$$
$$\begin{aligned} &\text{if } -3\sqrt{\lambda_1} \le b_1 < -0.5\sqrt{\lambda_1},\ h(b_1) = \Theta \\ &\text{if } -0.5\sqrt{\lambda_1} \le b_1 < 0.5\sqrt{\lambda_1},\ h(b_1) = \Gamma \\ &\text{if } \quad 0.5\sqrt{\lambda_1} \le b_1 < 3\sqrt{\lambda_1} \quad ,\ h(b_1) = \Lambda \end{aligned} \tag{4}$$
$$s = \exp\left[ -\min_{v=1,\ldots 5} \left( \left\| g_j - \bar{g}_{vj}(h(b_1)) \right\|_2^2 \right) \right] \tag{5}$$

where $\bar{g}_{vj}(h(b_1))$ represents the *v-th* cluster centroid computed from all the gradient profiles at $p_j$ in training subset $h(b_1)$. Function $h$ returns the pose subset based on the value of $b_1$.

# 4. Bayesian parameter update with statistical sampling

Statistical analysis of the positions of a number of landmarks located on training images provides a generic face template, average face $\bar{\mathbf{f}}$ and basis vectors $\Phi$, which account for the *generalities* of face variations in the training set. However, image *specificities* have to be taken into consideration to *update* the coefficient vector **b** such that the reconstructed shape best matches the image data. In addition to the coefficient vector **b**, we also estimate the global translation ($X$, $Y$) of the face shape. This leads to the parameter set $\theta = (X, Y, \mathbf{b})$ with $\mathbf{b}=[b_1, b_2, ..., b_t]^T$.

Initialize $\theta = (X, Y, \mathbf{b})$ as $\theta = (X_0, Y_0, 0, 0, 0)$

Loop until stop-criterion satisfied

1. Generate $H$ samples $\theta_i$, $i$=1,..., $H$, Eq. (7)
2. Compute likelihood $p(I \mid \theta)$
   - 2.1 Construct shape from the parameters, Eq.(8)
   - 2.2 Compute the likelihood for each sample, Eq. (11)
3. Resampling proportional to the likelihood

Figure 5: The proposed algorithm.

Given a new image, the parameters can be calculated by estimating the posterior density of $\theta$ in a Bayesian framework:

$$p(\theta \mid I) = p(I \mid \theta)\, p(\theta) / \, p(I) \propto p(I \mid \theta)\, p(\theta) \tag{6}$$

In our algorithm, the parameters are updated iteratively. At the beginning, the coordinates of the global translation, ($X$, $Y$), are initialized by face detection; and the elements of the vector **b** are initialized to all zeros. In general, the posterior probability density, $p(I|\theta)$, is highly non-linear and multi-modal. Thus no parametric form of $p(I|\theta)$ should be assumed. Thus, we use a set of weighted samples to approximate the posterior PDF iteratively. This leads to a CONDENSATION-like statistical sampling scheme. One advantage of statistical sampling is that it allows us to know the rough pose of the current shape instance, therefore we can compute pose-dependent likelihood (Sec. 4.2). Our work differs from prior CONDENSATION scheme [11] in that we introduced the pose-dependent likelihood model, which was found to account for a large portion of the improved performance. The proposed algorithm is shown in Fig. 5, where the stop criterion is currently set as a fixed number of iterations. When the algorithm terminates, the mean shape computed from the weighted samples is kept as the final model. In the following, we discuss each step in detail.

## 4.1. Generate random samples

Face detection and prior statistics on the suitable limits of $b_k$ effectively constrain the parameters to be within a certain region in a high-dimensional parameter space. This allows us to generate $H$ random samples that are uniformly distributed within a hyper-rectangle:

$$\left\{ \begin{aligned} \theta_i = (X_i, Y_i, \mathbf{b}_i) \Big| & X_i^{r-1} - \delta_X \le X_i^r \le X_i^{r-1} + \delta_X \\ & Y_i^{r-1} - \delta_Y \le Y_i^r \le Y_i^{r-1} + \delta_Y \\ & -3\sqrt{\lambda_k} \le b_{ik}^r \le 3\sqrt{\lambda_k}, k = 1, \ldots t. \end{aligned} \right\} \tag{7}$$

One could model $p(\theta)$ using a mixture of Gaussians with a kernel density estimate for each Gaussian component [9], or even learn a prior non-parametric density from the training set (which is only meaningful if the number of training images is large enough). However, in our current system, we found random sampling with uniform distribution works quite well.

## 4.2. Likelihood model $p(I|\theta)$

The usefulness of the generated samples needs to be evaluated based on the image measurement, i.e. how well the reconstructed face shape from a parameter sample fits the image data. The local image measurement given a face configuration is defined as the gradient profiles estimated along the line perpendicular to the model boundary through each model point, as in [19].

*Reconstruct shape from one parameter sample*. Given a parameter sample $\theta_i = (X_i, Y_i, \mathbf{b}_i)$, the corresponding face is reconstructed by

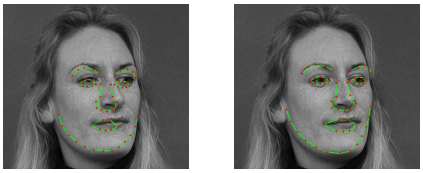
$$dx = X_i - X_a \qquad dy = Y_i - Y_a \tag{8.1}$$
$$\boldsymbol{f}_i = T_{dx,dy}(\bar{f}) + \Phi \mathbf{b}_i \tag{8.2}$$

In Eq. (8.1), we compute the offset between the face location ($X_i$, $Y_i$) of a model instance and the center of the bounding box that encompasses the landmarks of the *mean shape* $\bar{f}$ (computed from training images). Then in Eq. (8.2), the $i$-th shape instance is reconstructed by moving the mean shape to the current face location with transformation $T_{dx,dy}(\bar{f})$ and adding the intrinsic shape deformation $\Phi \mathbf{b}_i$. Function $T_{dx,dy}(\bar{f})$ performs a translation by ($dx$, $dy$).

*Likelihood model.* Let $g_j$ denote the gradient profile computed along the line perpendicular to the boundary of a shape instance $f_i$ through landmark point $p_j$. We define the likelihood as

$$p(I \mid \theta_i) = \eta \bullet p(g_1, \cdots, g_N \mid \theta_i) + (1-\eta) \bullet p(G(\theta_i)) \\ = \eta \bullet \left( \prod_{j=1}^{N} p(g_j \mid \theta_i) \right) + (1-\eta) \bullet p(G(\theta_i)) \quad (9)$$

The likelihood consists of two terms: *model-driven* part $p(g_1, ..., g_N \mid \theta_i)$, which is the joint probability of the gradient profiles at the $N$ local landmarks given the shape configuration $\theta_i$; And *data-driven* part $p(G(\theta_i))$ which measures how well the global face contour fits to the true face contour. The data-driven part increases the ability of the algorithm in generalizing to unseen images, and the facial features can be more accurately located than only using the model-driven part. As illustrated in Fig. 6.



(a) (b)

Figure 6: *k*-means clustering and data-driven part in the likelihood function increase the accuracy of face alignment. (a) Alignment result with neither technique. (b) Alignment result with both *k*-means clustering and data-driven likelihood.

The likelihood at each landmark, $p(g_j \mid \theta_i)$, is given by

$$p(g_j \mid \theta_i) = \exp\left[ -\min_{v=1,\ldots,5} \left( \left\| g_j - \bar{g}_{vj}(h(b_{i1})) \right\|_2^2 \right) \right] \quad (10)$$

where $\bar{g}_{vj}(h(b_{i1}))$ represents the *v-th* cluster centroid computed from all the gradient profiles at $p_j$ in training subset $h(b_{i1})$, as discussed in Sec. 3.2.

The data-driven part, $p(G(\theta_i))$, is computed as

$$p\left(G(\theta_i)\right) = \frac{\alpha}{\beta} \quad (11)$$

$\alpha$ denotes the total number of edge pixels encompassed by all the 3x3 windows centered at each point located on the facial feature contour that fall inside of the bounding box of the true face obtained by face detection. $\beta$ denotes the total number of edge pixels that fall inside of the bounding box of the true face obtained by face detection. $p(G(\theta_i))$ gives us the ratio of the model edge pixels to the true face edge pixels.

*Weighting differently for different landmarks.* We also observed that when the face turns left (right), the collection of landmarks located on the left (right) side of face will have smaller contribution to the likelihood computation. Suppose we know the rough orientation of a shape instance, the likelihood model can be further improved such that the image measurements of different set of landmarks are weighted differently. In our statistical sampling framework, it is easy to achieve that because $b_{i1}$ gives the rough orientation of a face instance. In particular, we partition the $N$ landmarks into three sets, $S_1$, $S_2$ and $S_3$, corresponding to the set of right-side, middle and left-side landmarks (Fig. 2). If $0.5\sqrt{\lambda_k} \le b_{ik} < 3\sqrt{\lambda_k}$, we deem the current face is turning right, then the landmarks in set $S_1$ will be assigned smaller weights than those in $S_2$ and $S_3$; If $-3\sqrt{\lambda_k} \le b_{ik} < -0.5\sqrt{\lambda_k}$, we deem the current face is turning left, then the landmarks in $S_3$ will be assigned smaller weights than those in $S_1$ and $S_2$. If $-0.5\sqrt{\lambda_k} \le b_{ik} < 0.5\sqrt{\lambda_k}$, all the three sets are viewed equally important. This yields the following enhanced likelihood model:

$$p(I \mid \theta_i) = \eta \bullet \exp\left[ -\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}^T \bullet \begin{pmatrix} \sum_{j \in s_1} \min_{v=1,\ldots,5} \left\| g_j(\boldsymbol{f_i}) - \bar{g}_{vj}(h(b_{i1})) \right\|_2^2 \\ \sum_{k \in s_2} \min_{v=1,\ldots,5} \left\| g_k(\boldsymbol{f_i}) - \bar{g}_{vk}(h(b_{i1})) \right\|_2^2 \\ \sum_{l \in s_3} \min_{v=1,\ldots,5} \left\| g_l(\boldsymbol{f_i}) - \bar{g}_{vl}(h(b_{i1})) \right\|_2^2 \end{pmatrix} \right] \\ + (1-\eta) \bullet p(G(\theta_i)) \quad (11)$$

In the re-sampling step, we multiply samples with large image likelihood and discard samples with small likelihood. In implementation, we also perform a local search for mean shapes of eyes, nose, mouth, and the jaw respectively, after the final re-sampling step.

### 4.3. Edge based enrichment

In the tactile representation, in addition to the contours of the major facial features like eyes, mouth, and nose, it is also essential to keep other informative edges in the final tactile rendering. To achieve this, we enrich the basic face model obtained from our face alignment algorithm with the edges detected by a Canny edge detector. An adaptive edge refining step is used to filter out the redundant edges. E.g., for eyes, eyebrows, jaw and nose, we keep their shape contours and the nearby edge segments; For mouth, we did not use the contour of the landmarks obtained from the face alignment algorithm; rather, in order to better reflect the expressions conveyed by the mouth such as smiling, we keep the original edges that are located inside a bounding box around the detected mouth landmarks.

### 4.4. 2.5-D tactile graphical representation

The edge-enriched portrait image can be transformed to tactile form by printing it out using an embosser or a thermoform machine. In this stage, we exploit the strength of the gradient to modulate the tactile patterns in generating the tactile graphics. For example, we use denser dot patterns for areas with strong gradients, thicker lines for major facial features, and thin lines for the secondary features including wrinkles, fine edges around the eyes and the mouth.

(a) (b) (c) (d) (e)

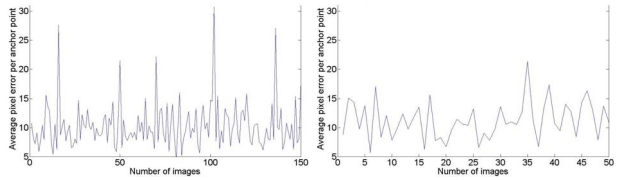
Figure 7: The entire tactile conversion process. First row shows original images, 2nd row shows the results of face alignment, 3rd row depicts the edge-enriched representation, 4th row illustrates the tactile printouts from a thermoform machine.

Figure 7 illustrates with examples the steps of the tactile conversion process. The first row lists 5 input images, the second row illustrates the face alignment results, the 3rd row is the edge-enriched face renderings, and the 4th row shows the pictures of the actual tactile printouts.

## 5. Experiment set-up, results and evaluation

### 5.1. Face alignment evaluation

We used two face image databases to evaluate the performance of the proposed face alignment algorithm. The first database, IMM [16], comprises 240 images of 40 different human faces. Each person has 6 poses: 4 full frontal poses with varying lighting, 1 pose turning approximately 30˚ to the right, and 1 pose turning approximately 30˚ to the left. We use 30 persons for training and the remaining 10 persons are used for testing. Fig. 8 shows some sample results of the alignment algorithm for different poses and lighting conditions. We also quantitatively analyzed the performance of the algorithm by comparing the estimated shape with the ground truth shape and plotting the error, as shown in Fig. 9. We can see that the average error per anchor point is about 10 pixels for both training and test images. Since the average height of face regions in the database is about 200 pixels, the error is only about 5% of the height of the face region, and thus can be deemed as small.




(a) (b)

Figure 9: (a) Average errors per anchor point of training images. (b) Average errors per anchor point of test images.

Our results are also comparable to what presented in [9] although it is difficult to make straightforward comparison since the databases are different. In our results, the percentages of images with average per anchor point error less than or equal to 13.3 pixels are 84% and 74% for the training set and the test set respectively. These are compared to 89.7% and 57.8% reported in [9] for group1 and group2 respectively, for images with errors less than or equal to 12 pixels cases. (The average height of the face region in the database used in [9] is about 180 pixels. Hence we compare our 13.3-pixel error against their 12-pixel error.)

In order to illustrate the robustness of the algorithm in fitting to novel images whose acquisition environment differs greatly from the IMM database, we independently captured a 30-person face database with varying lighting conditions, expressions and poses. The resolution of the images is also much lower than the first set. A few sample results are shown in Fig. 10, demonstrating the robust performance of the algorithm.

Figure 8: Face alignment results. The faces in the 1st row are frontal with some variations of expressions and lighting. The faces in the 2nd and 3rd row have off-plane rotations.

Figure 10: Face alignment results on the second data set.

## 5.2. Tactile representation evaluation

In order to evaluate the effectiveness of the tactile graphics, we did two sets of experiments. The objective of the first experiment is to evaluate whether the tactile portrait image can effectively convey the important and informative facial features. Six volunteers including 5 blind-folded sighted persons and 1 user who is blind participated in the experiment. Although the end user of the technology will be people with visual impairment, at this stage of study, to verify that the approach does maintain key "visual" features, it was found that using blind-folded sighted individuals for the evaluation is very helpful since they are able to compare what they feel by touch against what they see. (This is especially true for the second evaluation experiment.)

For each user, we asked him/her to touch the 5 tactile portraits presented in Fig. 7, last row, one by one. They need to answer the following 4 questions: (1) Can you recognize each face component including the mouth, eyes, eyebrows and nose? (2) Can you recognize the pose of the person, i.e. is he/she turning left or turning right? (3) Can you recognize the gender of the person? (4) Can you identify two images that represent the same person? The results were collected in Table 1, where (for example) 5/6 means 5 out of the 6 testers got that answer correctly and "Association" refers to Question 4.

Table 1: The resultant statistics obtained from the first experiment of tactile representation evaluation.

| Image | Left eye, eyebrow | Right eye, eyebrow | Nose | Mouth | Pose | Gender | Associ-ation |
|---|---|---|---|---|---|---|---|
| a | 6/6 | 5/6 | 6/6 | 6/6 | 6/6 | 5/6 | 5/6 |
| b | 6/6 | 6/6 | 5/6 | 5/6 | 6/6 | 6/6 | |
| c | 6/6 | 5/6 | 6/6 | 6/6 | 6/6 | 3/6 | |
| d | 4/6 | 4/6 | 4/6 | 5/6 | 5/6 | 3/6 | |
| e | 5/6 | 5/6 | 4/6 | 5/6 | 5/6 | 5/6 | |

The chart indicates a few interesting facts. First, the user who is blind was much faster than other blind-folded sighted persons in interpreting the results; she gave all correct answers except the gender of image *e*. She could even correctly recognize facial expressions (smiling, frown) based on the shape of mouth (open/close, wide/thin). She started to identify the face components mainly based on the shape of the mouth and the nose, whereas the blind-folded sighted person relied mainly on the hair and the global contour to get started. Second, gender recognition is a little harder than other tasks because it is difficult to differentiate hair from shoulder, if purely relying on the haptic sense. Third, the length of the right/left eyebrow and the presence/absence of ears play critical role in delivering the pose information (turning left/right) since it is usually true that when the face turns right, the length of left eyebrow will be longer and the ear will become visible on the image. Fourth, pose recognition has direct effect on the identification of nose and mouth, because if the pose is not correctly recognized, then the user tends to explore the wrong parts of the image for the nose/mouth. Because of the importance of correct pose recognition, we can print in Braille the pose of the face on

the tactile graphics; this will significantly help the understanding of tactile portraits.

The objective of the second experiment is to analyze whether the system is able to retain the distinctive characteristics of the facial features. Five blind-folded sighted persons participated in the experiments, and two tactile images (Fig. 11) are used for their exploration. The participants were asked to give the identity of each person on the two tactile images, chosen from 5 persons whom they all visually know very well. For Example, we asked the participants: "Can you tell who this person is, chosen from Cindy, Troy, Jessie, Michael, and Daniel?" The results were very encouraging: all of the 5 users were able to correctly recognize the identity of the persons on the two images, which shows that the converted tactile representation does a good job in retaining the distinctive visual characteristics.

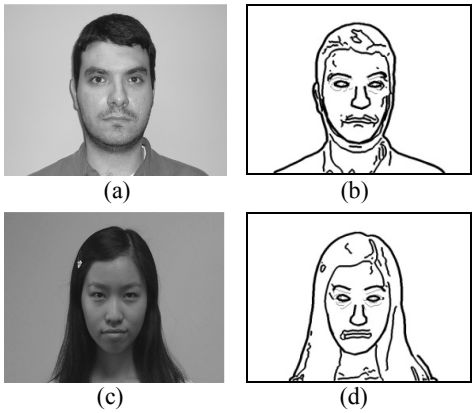



Figure 11: The two orignal images (a) and (c) and the corresponding tactile forms (c) and (d) used in the second experiment for tactile graphics evaluation.

## 6. Conclusion and Future Work

This paper proposed a novel solution for automatic conversion of digital portrait images into their tactile forms. Using a Bayesian inference framework, the approach utilizes a statistical-sampling-based algorithm for aligning a prior face model with a given image, which allows us to avoid getting stuck in the local minima. The likelihood model in our approach includes a model-driven part and data-driven part, which increases the robustness of the algorithm when generalizing to unseen images. A pose-dependent likelihood model was proposed to facilitate the match between the data and the model. The aligned contours of the facial components are enriched with local informative edge segments to improve its capability in retaining the most distinctive characteristics of the faces. Experiments and evaluation on both face alignment and tactile conversion show that the proposed approach is very effective, suggesting this is a promising approach to the challenging problem. A full-scale evaluation with a large number of blind users with diverse educational/training background, level of exposure to tactile graphics, level of visual memory, etc. will be among our future tasks.

## References

[1] Satoshi Ina. Presentation of images for the blind. *ACM SIGCAPH Computers and the Physically Handicapped*, 56: 10-16, 1996.

[2] T. Way, K. Barner. Automatic visual to tactile translation, part I: human factors, access methods and image manipulation. IEEE Transactions on Rehabilitation Engineering, vol. 5, pp. 81-94, Mar. 1997.

[3] R.E. Ladner, M.Y. Ivory, R. Rao, S. Burgstahler, D. Comden, S. Hahn, et al. Automating tactile graphics translation. The 7th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '05), pp. 50-57.

[4] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham. Active shape models – their training and application. Computer Vision and Image Understanding, 61(1):38-59, 1995.

[5] http://dots.physics.orst.edu.

[6] Tactile Graphics Project at University of Washington: http://tactilegraphics.cs.washington.edu.

[7] Y. Zhou, L. Gu, and H. Zhang. Bayesian tangent shape model: estimating shape and pose parameters via Bayesian inference. CVPR (1), pp. 109-116, June 2003.

[8] F. Jiao, S. Li, H.-Y. Shum, and D. Schuurmans. Face alignment using statistical models and wavelet features. In Proc. CVPR (1), pp. 321–327, 2003.

[9] Y. Huang, Q. Liu, D. Metaxas1. A component based deformable model for generalized face alignment. In Proc. ICCV, 2007.

[10] X. Liu. Generic face alignment using boosted appearance model. In Proc. CVPR, pp.1-8, 2007.

[11] J. Tu, Z. Zhang, Z. Zeng, T. Huang. Face localization via hierarchical CONDENSATION with Fisher boosting feature selection. CVPR (2), pp. 719-724, 2004.

[12] J. Coughlan and S. Ferreira. Finding deformable shapes using loopy belief propagation. In Proc. ECCV, 2002.

[13] L. Liang, F. Wen, Y.-Q. Xu, X. Tang, and H.-Y. Shum. Accurate face alignment using shape constrained Markov network. In Proc. CVPR (1), pp. 1313-1319, 2006.

[14] N. D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. In Journal of Machine Learning Research 6, pp 1783–1816, 2005.

[15] C. Hu, R. Feris, and M. Turk. Active wavelet networks for face alignment. In Proc. 14th British Machine Vision Conference, Norwich, UK, 2003.

[16] M. B. Stegmann, Bjarne K. Ersbøll, and Rasmus Larsen. FAME - a flexible appearance modeling environment. IEEE Trans. On Medical Imaging, 22(10): 1319-1331, 2003.

[17] J. C. Gower. Generalized procrustes analysis. Psychometrika, 40:33-50, 1975.

[18] G. J. Edwards, A. Lanitis, C. J. Taylor, T. F. Cootes. Statistical models of face images – improving specificity. Image and Vision Computing, vol. 16, pp. 203-211, 1998.

[19] T. F. Cootes and C. J. Taylor. Statistical Models of Appearance for Computer Vision, pp. 12-28, 37-43, 2004.