Abstract

We present a data-driven method that generates new human faces from a sparse base of geometric information. Given a multiple sets of geometric facial data that singularly span variations in viseme, type, and expression, we construct faces that arbitrary combine these three types of variation to generate new face data. Our method relies on user-driven tagging of each existing face model to guide an algorithm which learns the space of variation in all faces. It links abstract bases of geometric variation to continuous, intuitive, and human readable attributes which correspond to viseme, type, and expression contribution. The output of our algorithm is a representation of a human face that can be continuously posed by the modification of these same attributes to create arbitrary new faces. We capture complex global facial transformations associated with these three types of attributes and allow robust control with a small set of continuous and concrete parameters.

1. Introduction

The process of generating facial models and various poses of these models is a necessary part of most present-day movies, which are heavy in computer-assisted effects, and almost required for any interactive game which features humans as a primary character. The generation of this face data can be approached in ways varying from pure computation to pure data acquisition. Computational methods employ a representation of the underlying physical components of a human face, such as bone, muscle, fat, and skin, which then must interact in a physical simulation and generate deformation information for a three dimensional mesh. These models are computationally expensive to modify, require extensive experimentation with physical parameters, and if not accurately represented, can result in facial deformations that are unnatural or unattractive. Purely data-based models rely on capturing real-world facial data from range scans or motion capture information and then associating each piece of data with a set of semantically meaningful parameters for grouping and retrieval. These types of face models are based on real world data, and are therefore, by definition, realistic. Unfortunately, the variation of face poses are limited by the data set, and the process of acquiring new face data for specific expressions or face types requires the recapture of new face data, which can be an expensive and time-consuming process.

Our method is a hybrid approach, which combines a relatively small set of real world facial data with an algorithm that learns the underlying variations in this geometric information automatically. Given a sparse data set that spans variation in viseme, face type, and expression, we are able to generate new faces that exhibit combinations of these attributes, and were never part of the original data set. A diagram of the contents of our data set is displayed in Figure 1. We rely on user-assisted categorization of our sparse data set to associate each piece of face data with various attribute contributions, and then use this categorization data as a guide for binding abstract variation to concrete parameters. Our contribution is a method that takes the complex, subtle, and often subjective qualities associated with visemes, expressions, and face types, and correlates them to the contributions of the geometric bases that make up the entire space of variation in an existing set of face data, in order to facilitate the creation of entirely new face poses.

2. Related Work

Facial motion retargeting is often used to generate realistic expressions in synthetic 3D models. This involves the transfer of facial expression from a real human actor to a computer generated model. Guenter [Guenter98] explores a system that synthesizes a 3D model from images and then retargets an actor’s facial expressions to it. Curio [Curio06] retargets facial animation by decomposing the original motion capture data into small meaningful motions and then reapplying them on a morphable 3D model. While these papers retarget facial motion to new 3D models, they do not attempt to modify the original mouth movements or morph expressions into new expressions. They only generate facial motions that
were contained in their original data set.

Figure 1. Our algorithm makes use of a sparse set of face data that spans variation in viseme, expression, and face type. We generate new faces that span combinations of each of these “axes” of variation.

Other works have explored facial motion modification in order to create new motions not found in the original image or video. Pighin [Pighin98][Pighin99] creates new facial expressions by morphing two of the original motion capture expressions. Blanz [Blanz03] interpolates between expressions and visemes by establishing a common representation of different faces in a vector space of 3D shapes and textures. These works are able to morph along one dimension of variation, either the expression or type axis. However they are unable to morph in multiple dimensions. They do not take data from each of the type, expression, and viseme axes and generate new motions.

More recent works attempt to synthesize new facial animations in multiple dimensions by linking motion capture sequences that express combinations of type, expression, and visemes. [Chuang02][Chuang05] establishes a bilinear model of the two main factors influencing the face during speech, the speech viseme and facial expression. These factors are combined to create new speech animations with varying expressions and visemes. Vlasic [Vlasic05] pushes further with a multi-linear model of three shape factors (i.e., expression, viseme, and type). However both Chuang and Vlasic are limited to the data combinations captured in their source data, which must be a full Cartesian product of the desired facial types, expressions, and visemes. They are not able to generate new combinations of type, expressions, and visemes that are not included in this data set. They cannot handle the level of sparsity present in our own data set.

The synthesis of new attribute combinations from a sparse data set is addressed by Allen [Allen04] with respect to human body shapes. They use principle component analysis and multi-attribute regression to learn how each component relates to body shape factors such as height and shoulder width. With the computed parameters, they are able to synthesize new body shapes with any desired attributes. Our approach is most similar to this work, except we apply it to facial poses which have subjective attributes. Whereas Allen uses body shape attributes with clearly objective attributes such as bone length, our work requires human input to determine subjective facial attributes such as masculinity and sadness.

None of these previous works synthesize new combinations of facial expressions, visemes, and types from a sparse data set. They do not morph between all variations of the three components of face shape. They do not morph multiple expressions into a new expression, multiple identities into a new identity, and combine them with a viseme to synthesize a completely new facial motion. Our paper explores a system that allows seamless control between multiple expressions, visemes, and types.

Figure 2. All face models in our sparse data set are tagged by a human with attribute values for type, expression and viseme attributes. Each tag, a continuous value from zero to one, represents the apparent strength of an attribute in a single face model.

3. Methods

Our algorithm is separated into three primary stages: tagging, mesh analysis, and generation. During the tagging
step, a user is presented with each model from the original facial data set, and he or she judges the contributions of various attributes for that specific model. Each attribute contribution is a continuous value from zero to one, represented by a slider in our user interface. The three categories of attributes are face types, expressions, and visemes, as seen in Figure 2. Face types can be arbitrary qualities that vary from high level qualities such as femininity or weight, to specific features such as chin definition. Expression attributes are the contribution of a basic set of emotions, including joy, sadness, anger, fear, disgust and surprise. Our viseme attributes are derived from the known set of visemes and their associated phonemes, studied in previous research [Ezzat00].

Once every face model has been tagged with attribute contributions by a user, we then analyze how these contributions correlate directly to the principle components of geometric variation that exist throughout our entire data set. We assume that all faces in our data set are in one-to-one vertex correspondence, which is achieved using previously discussed methods [Allen04]. We first generate a reference face model for set of face data, which is an average of all of our existing face models. We then represent each vertex of each unique face in that set as an offset vector from the position of the same vertex in the reference face. We perform principle component analysis on all of these offsets for every face model in that specific data set, and save all of the relevant bases of variation, or eigenvectors. Then, for each face model, we store the contribution, or weight, of each eigenvector, so that applying them in linear combination to the vertices in the reference face results in a shape that approaches the actual location on each original face model. Each face in a single data set is therefore constructed as follows:

$$f_i = a + \sum_{j=1}^{N} \text{weight}_j v_j$$

In equation 1, \(f_i\) is a vector of locations for all vertices in face model \(i\), and vector \(\text{weight}_j\) is the contribution of the common principle eigenvector \(v_j\) specific to face \(i\). As a result, each of the faces in our data set is represented by \(N\) scalar values corresponding to a common set of \(N\) eigenvectors, where \(N\) is the number of face models analyzed. The eigenvectors and weight vectors each contain \(M\) elements, where \(M\) is number of vertices in each face model.

We next compare these weight vectors to the user-submitted tagging information, and learn a relationship between the contributions of each eigenvector, or basis of geometric variation, and the strength of every arbitrary facial attribute throughout the data set. We perform this regression across all attributes and weights in the following way:

$$M = WA^{-1}$$

Here, \(M\) is an unknown transformation matrix, \(W\) is a matrix in which each column is a single weight vector for a single face model, as calculated in equation 1, and \(A\) is a matrix in which each column is the set of attribute ratings for every face, in the same order as the columns in matrix \(W\). This correlation process involves a least best-squares fit using a matrix pseudo-inversion, and is the same method used in previous work with body shape deformation [Allen04]. The above Principle Component Analysis and transformation matrix derivation step is performed three times independently, once for each separate set of faces and attributes for types, expressions, and visemes. When this stage is complete, we have solved for three relationship matrices, \(M_{\text{type}}, M_{\text{expression}}, \text{and } M_{\text{viseme}}\), that respectively correlate high level type, expression, and viseme attribute values to underlying abstract geometric variation in each set of face models.

The final step is the generation of new faces. With the data gathered from the tagging and analysis phases, we can generate a multitude of continuously variable and arbitrary new faces by modifying the small set of type, expression, viseme attributes used in the first step of our algorithm. First, we acquire a vector of desired attributes from a user, and transform them into a set of eigenvector weights, using the same matrices derived in the previous analysis step. The following equation demonstrates the relationship between these desired attributes and the generated weights of each principle component:

$$w_{\text{generated}} = Mw_{\text{desired}}$$

In this equation, \(w\) is a vector of new weights for each principle eigenvector, and \(a\) is a vector of user-inputted, desired attribute values for a new face. This calculation is performed three times with the three sets of requested attributes relating to type, expression, and viseme, to generate three new sets of weights, \(w_{\text{type}}, w_{\text{expression}}, \text{and } w_{\text{viseme}}\). The final step, in which we generate new geometry, occurs as follows:

$$\text{newface} = \sum_{i=1}^{A} w_{\text{type}}^i v_{\text{type}}^i + \sum_{j=1}^{B} w_{\text{expression}}^j v_{\text{expression}}^j + \sum_{k=1}^{C} w_{\text{viseme}}^k v_{\text{viseme}}^k$$

In equation 4, \(\text{newface}\) is a vector containing the final locations of all vertices in a generated face. Each summation represents the addition of type, expression, and
viseme vertex offsets to the average type face locations, morphing the final face into a shape that reflects the original requested attributes.

Figure 3. This figure shows the results of applying attributes over the space of faces. Each of the first three rows shows examples along a single axis of variation. The last row shows examples along all three axes of variation.

4. Results

With our technique, the user is free to explore the entire multi-dimensional space of variation that is possible amongst the input data set. We can modify the contribution of any facial type, expression, or viseme and create reasonable human faces. Our source data set consists of 459 human faces with 23,728 vertices each. Each face was tagged with the contributions of five facial types, six expressions, and five viseme attributes.

Example results of the generation process of our algorithm are displayed in Figure 3. Our generated type faces are realistic, because our face data set includes 75 different faces with neutral expressions. Our generated expression faces are highly realistic, because our data set includes 384 faces of a single actor performing various expressions. Our generated viseme faces are not very realistic and show minor expression changes, because we reuse the 384 expression faces as our viseme faces. Ideally we should have a set of faces showing pure visemes without any expression.

The appearance of occasional artifacts is a limitation of our algorithm. This is most apparent, for example, with an extremely disgusted expression. The lower and upper lips interpenetrate because the emotions expressed in our expression-based data set represent facial variation that is specific only to a single face size and shape. Our algorithm does not model physically-correct facial movement, or attempt to alter expression offsets to compensate for the size and structure differences in various face types. Other artifacts can also occur, such as unexpected eye changes when adjusting the chin or nose types. This is due to our simple global geometry analysis method, which does not isolate desired geometric variation, as well as inherent imperfections in the human tagging step of our method.

Another limitation is the creation of unrealistic facial poses. When certain expressions and visemes are combined, facial features appear unnaturally stretched or squished. When combining a joy expression with the /aa/ viseme, the mouth is opened too far and the cheeks are pushed into the eyes. This is due to the fact our expression data set was not ideal, and contained mouth movement that was not categorized as variation in visemes. As a result, the viseme offsets can unnaturally elongate the existing mouth movement embedded in our expression offsets.

5. Conclusion

We have shown how to take a sparse set of input facial data that spans a set of variations, and arbitrarily combine these variations to generate new facial data that did not previously exist in the input data set. We developed an algorithm that uses principal component analysis to learn variation, and linear correlation techniques to attach this variation to concrete parameters. Finally, our method uses

---

<table>
<thead>
<tr>
<th>Type Variation</th>
<th>Expression Variation</th>
<th>Viseme Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male With Strong Chin</td>
<td>Anger And Disgust</td>
<td>/æ/</td>
</tr>
<tr>
<td>Female With Weak Chin</td>
<td>Sadness And Fear</td>
<td>/a:/</td>
</tr>
<tr>
<td>Male, Light Weight, Fear, /p/</td>
<td>Female, Sharp Nose, Joy, /uh/</td>
<td>Male, Light Weight, Fear, Surprise, /p/</td>
</tr>
</tbody>
</table>

Type, Expression, And Viseme Variation

Female, Sharp Nose, Joy, /uh/ Male, Light Weight, Fear, Surprise, /p/
this analysis to facilitate the generation of faces with these same high level controls. Ideally, an animation system or interactive rendering engine can make use of this high level control to produce emotionally rich, realistic, and expressive motion without the complexities of direct geometry manipulation. Our method is ideal for generating realistic faces, because it leverages the use of a sparse and manageable set of real-world data, and at the same time is able to capture and intuitively represent the complex interactions between face types, expressions, and visemes simultaneously throughout an entire human face.

References


