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. While these types of face models are based on real world data, and are therefore, by definition, realistic, the amount of data required to span a complete set of variations in face poses and types can result in excessively large databases of information which are unwieldy to store. In addition, the process of acquiring new faces for new applications requires recapturing 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

Guenter [Guenter98] and Pighin [Pighin98] describe several techniques for creating photo-realistic 3D face models from images. By fitting a generic 3D mesh to the original face, they approximate a synthetic 3D model that looks like their original image. And by extracting facial expressions and textures from the original images, they transfer the facial animation to the synthetic model. However they only transfer the expressions to a 3D model of the original face, and they do not retarget the original facial expressions to new 3D models.

Other papers describe methods that retarget facial
motion to new 3D models. Blanz [Blanz03] transfers mouth speech movements and expressions to other models by establishing a common representation of different faces in a vector space of 3D shapes and textures. Curio [Curio06] retargets facial animation by decomposing the original motion captured data into small meaningful motions (Action Units) based on the Facial Action Coding System. These motions are replayed on a morphable 3D model that is created from registered 3D motion captured and scanned data. 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.

More recent works attempt to synthesize and retarget new facial motions that were not explicitly captured in the original motion data. Pighin [Pighin98] creates new facial expressions by morphing two of the original captured expressions. Their 3D shape morphing technique creates natural animations as the face changes from one expression to another. However, they do not consider facial types and visemes. Chuang [Chuang02][Chuang05] explores how expressions are modified during continuous speech. They establish 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. However, they are limited to the expression and viseme combinations captured in their source data, which must be a full Cartesian product of the desired facial expressions and visemes. They are not able to generate new combinations of expressions and visemes that are not included in this data set.

Vlasic [Vlasic05] pushes the synthesis of new facial motions further with a multi-linear model of three shape factors (i.e., expression, viseme, and type) and retargets the new animations to other models. Their multi-linear model also requires a full Cartesian product of the three shape factors, and they are able to infer a limited amount of missing face data, but, due to limitations in their imputation process, it 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 also 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 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.

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. Face types can be arbitrary qualities that vary from high level qualities such as maleness or femaleness, to specific features such as nose size. Expression attributes are the contribution of a basic set of emotions, including happiness, sadness, anger, fear, 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 in our sparse data set 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, which is an average of all of our existing face models. We then represent each vertex of each unique face model 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 our data set, and save the first 30 primary bases of variation, or eigenvectors. Then, for each face model, we solve for the best weight for each of these eigenvectors, 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 is therefore constructed as follows:

\[ f_i = a + \sum_{j=1}^{30} w_j v_j \]

In this equation, \( f_i \) represents the locations of all vertices in original face \( i \), \( a \) is the vector of all vertex positions in the reference face, \( w_j \) is the weight of principle component \( j \) for face \( i \), and \( v_j \) is vector of principle component \( j \). As a result, each of the faces in our data set is represented by 30 scalar values corresponding to a common set of 30 vectors, each of which contains \( N \) elements, where \( N \) is number of vertices in each face model.

We then compare these associated weights to the user-submitted tagging information, and learn a relationship between the contribution of each principle component, 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 the set of 30 weights representing a single face in our data set, and \( A \) is a matrix in which each column is the contribution of each possible attribute for that same face. 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]. When this stage is complete, we have determined a relationship matrix \( M \) between all concrete expression, viseme, and type attribute values and the underlying abstract spatial variations in the entire data set.

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. Starting with the reference face, each attribute is adjusted continuously between the values of zero and one. The following equation demonstrates the relationship between these desired attributes and the generated weights of each principle component:

\[ w_{generated} = Ma_{desired} \]

In this equation, \( w \) is a vector of new scalar weights for each principle component, and \( a \) is a vector of user-inputted desired scalar attribute values for a new face. As each attribute value changes, the contribution of each of the 24 principle components changes according to the precomputed relationship matrix from the previous analysis step in our algorithm. A new face is generated by offsetting each vertex in the reference face according to the weighted contribution of each principle component.

4. Results

Example results of the generation process of our algorithm are displayed in Figure 2. 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 viseme, expression, or face type and create reasonable human faces that correspond to the underlying geometric variation present in the original set of face data. Our source data set consists of 300 human faces with 17,000 vertices each, tagged with the contributions of 5 expression, 5 face type, and 12 viseme attributes.

A side effect of our algorithm is the potential generation of uninteresting or neutral faces as a result of conflicting attributes. When two or more attributes have competing transformation effects, and are applied with equal weight, the resultant faces may exhibit little facial movement, or show very little variation from the reference face. However, because we capture variation throughout the entire model, it is unlikely that any two attributes will completely cancel each other out and have zero effect on any part of the newly generated pose. In addition, the degree to which this problem is present depends on the variation in the underlying original data set, and accordingly, the distribution of the values associated with meshes during the tagging phase of our algorithm.
Figure 2. This figure shows the results of applying attributes over the space of faces. Images 1 and 2 show the variation from male to female. Variation between sadness and surprise is displayed in images 3 and 4. Images 5 and 6 show the visemes ‘uh’ and ‘w’ applied to a face.

5. Conclusion

We have shown how to take a sparse set of input facial data that spans a set of variations, and arbitrary 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 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 real-world data while requiring a sparse and manageable set of initial data, and at the same time is able to capture and represent the complex interactions between emotions, face types, and visemes simultaneously throughout an entire human face.

References


