Geração de poses de faces utilizando Active Appearance Model

Tupã Negreiros[1], Marcos R. P. Barretto[2], Jun Okamoto[3]

[1, 2, 3] Escola Politécnica da Universidade de São Paulo (POLI/USP)
Caixa Postal 61548 – CEP 05508-900 – São Paulo – SP – Brasil

tupa.negreiros@gmail.com, mrpbarre@usp.br, jokamoto@usp.br

Abstract. *Virtual presence is more and more important in communications and entertainment. Videocalls are possible nowadays; avatars are commonplace in computer games. This paper try to contribute to demonstrate the use the Active Appearance Model (AAM) technique to synthetize faces poses of persons. Results allow to foresee the use of AAM to add someone's face to a computer game, making interactivity even greater.*

Resumo. *A virtualização da presença é cada vez mais importante na comunicação entre pessoas e na indústria de entretenimento. As videochamadas já são uma realidade, bem como se encontram avatares em jogos de computador. Neste artigo, busca-se demonstrar o uso da técnica Active Appearance Model (AAM) para gerar poses de faces de pessoas. Os resultados permitem antever a aplicação da técnica em situações como a utilização da face de uma pessoa em jogos de computador, tornando a interatividade ainda maior.*

1 Introdução

Um avatar pode ser usado como uma representação para ambientes virtuais como jogos ou redes sociais, bem como em aplicações de videoconferência, pois utiliza menor banda de transmissão de dados do que um vídeo ao vivo. Avatares podem ser utilizados para tornar mais natural a interação homem-máquina.

É importante que o avatar seja capaz de gerar detalhes sutis da face, que são indicativos de expressões de emoções e da personalidade [Lee, Elgammal e Metaxas 2006].

O método *Active Appearance Model* (AAM) permite a criação de um modelo para geração de poses da face com uma quantidade relativamente pequena de imagens e sem o uso de sensores ou câmeras especiais. Requer, entretanto, o treinamento manual ou semiautomático de um conjunto de imagens.

2 Metodologia

O *Active Appearance Model* (AAM) tem esse nome pois incorpora parâmetros de forma e níveis em escala de cinza [Edwards, Taylor e Cootes 1998]. Pode ser utilizado como um algoritmo para reconhecimento de imagens [Edwards, Taylor e Cootes 1998], entretanto, neste trabalho, será utilizado para gerar poses de faces a partir do modelo criado. A geração é realizada variando parâmetros internos deste modelo.

[Lee, Elgammal e Metaxas 2006] utilizam métodos 3D criando uma estrutura e aplicando uma textura 2D. Ou ainda é possível utilizar uma variação do AAM para três dimensões chamado 3D *Morphable Model* (3DMM) [Jianglong Chang, Ying Zheng e Zengfu Wang 2007].

Outros autores [Abboud, Davoine e Mo Dang 2003A] e [Abboud, Davoine e Mo Dang 2003B] já sugerem utilizar o AAM para síntese de faces, mas com intenção de gerar seis expressões de emoções: alegria, tristeza, raiva, desgosto, medo e surpresa.

2.1 Treinamento

O modelo deve ser treinado com uma quantidade suficiente de imagens, não há *a priori* como determinar a quantidade de imagens para desejada qualidade final. Cada imagem precisa ser marcada com pontos representando as características buscadas, manualmente ou de forma semiautomática [Jianglong Chang, Ying Zheng e Zengfu Wang 2007].

O modelo é representado colocando o conjunto de pontos de cada imagem i em um vetor x_i, contendo n_0 pontos de treinamento, da seguinte forma [Cootes et al. 1995]:

$$x_i = \left(x_{i1}, y_{i1}, x_{i2}, y_{i2}, \ldots x_{in_0}, y_{in_0}\right) \qquad (1)$$

Considera-se uma rotação θ e escala s e uma translação t entre a forma de cada imagem em relação à média, de forma a alinhá-las, para isso se deve reduzir a soma ponderada:

$$E_j = \left(x_i - M\left(s_j, \theta_j\right)\left[x_j\right] - t_j\right)^T W \left(x_i - M\left(s_j, \theta_j\right)\left[x_j\right] - t_j\right) \quad (2)$$

Onde:

$$M(s,\theta)\begin{bmatrix} x_{jk} \\ y_{jk} \end{bmatrix} = \begin{bmatrix} s\cos\theta * x_{jk} - s\sin\theta * y_{jk} \\ s\sin\theta * x_{jk} + s\cos\theta * y_{jk} \end{bmatrix} \quad (3)$$

$$t_j = \left(t_{xj}, t_{yj}, \ldots t_{xj}, t_{yj}\right) \text{ com } 2n_0 \text{ elementos} \quad (4)$$

W é uma matriz diagonal de pesos para cada ponto, peso este inversamente proporcional a soma das variâncias de cada ponto em relação aos demais.

Como a quantidade de pontos demarcados é em geral pequena para representar uma imagem, executa-se uma triangulação dos pontos descrita por [Cootes e Taylor 1994] gerando tantos pontos quanto se queira. Como exemplo, tendo uma figura como um olho, com n_0 pontos demarcados. A partir desses pontos, é feita uma triangulação, determinando segmentos (Figura 1).

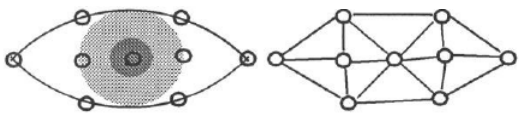

Figura 1 – Olho demarcado e triangulação [Cootes e Taylor 1994]

Então se dividem os segmentos na metade, obtendo-se novos pontos nos pontos médios. Pode-se triangular novamente (Figura 2), e repetidamente podem-se obter quantos pontos se desejar.

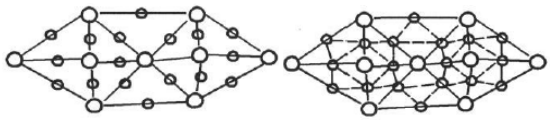

Figura 2 – Divisão dos segmentos e nova triangulação [Cootes e Taylor 1994]

Não é necessário alinhá-los novamente minimizando a equação (2), pois a triangulação foi executada com os pontos já alinhados.

Preenchendo o vetor x_i com os novos pontos na forma da equação (1), só que agora com n pontos, e aplicando uma análise de componentes principais (PCA) aos dados, chega-se que qualquer conjunto pode ser aproximado por [Cootes, Edwards e Taylor 1998]:

$$x_i \cong \bar{x} + P_s b_{si} \qquad (5)$$

Onde x_i é um conjunto de pontos de uma imagem, $\bar{x}$ é a média dos conjuntos de pontos, P_s é uma matriz com as variações na forma ortogonal e b_{si} é um conjunto de parâmetros.

Destes n pontos, cada qual tem um valor de intensidade em nível de cinza I em cada imagem. Como não necessariamente os pontos intermediários têm coordenadas inteiras, deve-se fazer uma interpolação bilinear para obter a intensidade em tal ponto. Alinhando as coordenadas de cada ponto, para cada imagem tem que:

$$g_i = (I_{i1}, I_{i2}, \ldots I_{in}) \qquad (6)$$

A fim de minimizar o efeito da luminosidade, podem-se normalizar os dados aplicando uma escala α e um offset β à informação de escala de cinza de g_i.

$$g_i = (g_i' - \beta_i \cdot 1) / \alpha_i \qquad (7)$$

Em que g_i' é o valor original do vetor g_i. Escolhe-se α_i e β_i da seguinte forma:

$$\alpha_i = g_i \cdot \bar{g} \text{ e } \beta_i = (g_i \cdot 1) / n \qquad (8)$$

Note-se que é um processo iterativo. Aplicando uma análise de componentes principais (PCA) à equação (6) de maneira análoga à equação (5), tem-se a aproximação:

$$g_i \cong \bar{g} + P_g b_{gi} \qquad (9)$$

Em que g_i é um vetor normalizado de nível de cinza de uma imagem, $\bar{g}$ é a média desses vetores de todas as imagens de treinamento, P_g é uma matriz com as variações na forma ortogonal e b_{gi} é um conjunto de parâmetros de nível de cinza.

Desse modo, a forma e a aparência de qualquer imagem podem ser descritas pelos vetores b_{si} e b_{gi}. Pode-se então definir um vetor b_i' concatenando os dois, de forma que:

$$b_i' = \begin{bmatrix} b_{si} \\ b_{gi} \end{bmatrix} = \begin{bmatrix} P_s^T (x_i - \bar{x}) \\ P_g^T (g_i - \bar{g}) \end{bmatrix} \qquad (10)$$

Pode-se ainda adotar uma matriz diagonal W_s com pesos para cada parâmetro de forma, de modo a permitir diferenças entre as unidades da forma com o nível de cinza:

$$b_i = \begin{bmatrix} W_s b_{si} \\ b_{gi} \end{bmatrix} = \begin{bmatrix} W_s P_s^T (x_i - \bar{x}) \\ P_g^T (g_i - \bar{g}) \end{bmatrix} \qquad (11)$$

Aplicando novamente o modelo da análise de componentes principais:

$$b_i = Q c_i \qquad (12)$$

Onde Q representa os autovalores e c_i os parâmetros de aparência, que incorporam tanto a forma como os níveis de cinza. Sendo que os parâmetros de forma e os de nível de cinza têm média zero, e c_i também tem.

Pela linearidade do modelo, pode-se expressar a forma e o nível de cinza como funções diretas de c_i:

$$x_i = \bar{x} + P_s W_s Q_s c_i \text{ e } g_i = \bar{g} + P_g Q_g c_i \qquad (13)$$

$$\text{onde } Q = \begin{bmatrix} Q_s \\ Q_g \end{bmatrix} \qquad (14)$$

2.2 Variação do modelo

Dessa forma, uma imagem pode ser sintetizada para um dado c_i por meio da variação da forma x_i e do nível de cinza sem-forma g_i.

Na Figura 3 é possível ver como a variação dos parâmetros c_i dos autovalores e autovetores mais importantes modificam a imagem sintetizada, com variações de ±3 vezes o desvio padrão, sendo as imagens da coluna central o modelo sem variação dos parâmetros.

Figura 3 – Quatro primeiros modos de variação do modelo [Cootes et al. 1995]

3 Resultados

O AAM busca maior variação do modelo, com isso geralmente ele é treinado com grande quantidade de pessoas diferentes, homens e mulheres, com maquiagem e sem, com pelos faciais e sem, com óculos ou não.

Já a proposta deste trabalho é treinar um modelo de AAM com faces de uma mesma pessoa. Com isso se desejam obter variações de poses e expressões faciais da pessoa.

3.1 Dados utilizados

Foram treinadas trinta imagens da face do presidente americano Barack Obama. Em cada imagem foram marcados manualmente 41 pontos. Pontos foram marcados ao redor da face para delimitar a região; e maior quantidade de pontos foi colocada nas regiões dos olhos e sobrancelhas e na região da boca, para melhor visualização das expressões faciais. Também foram colocados pontos nas narinas, pela facilidade de rastreamento, pois são pontos escuros no centro da face [Petajan e Graf 1996].

Na Figura 4 pode-se ver uma imagem com os 41 pontos demarcados.

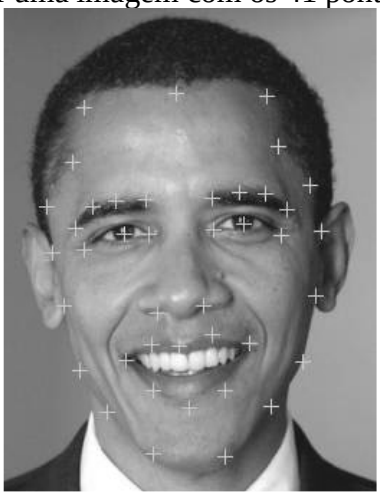

Figura 4 – Imagem de Barack Obama com 41 pontos

3.2 Criação do modelo

No processo de triangulação descrito no capítulo 2.1, a partir dos originais 41 pontos após o processo obtiveram-se 4.987 pontos.

Como componentes principais de forma descritos pela equação (5), apenas dois componentes representam 92,69% das variações. Já para os componentes principais de escala de cinza, da equação (9), dezesseis componentes representam 90,94% das variações.

Ao final do processo de PCA, dado pela equação (12), onze componentes de aparência (forma e escala de cinza) representam 91,95% das variações.

Aplicando o modelo nas imagens utilizadas, pode-se obter o desvio padrão do modelo em cada componente dos vetores b_{si}, b_{gi} e c_i, tais valores são apresentados nas Tabelas 1, 2 e 3 respectivamente. São valores importantes para a geração das imagens de variações do modelo mais a frente.

Tabela 1 – Desvio padrão do modelo de forma

j	**Desvio Padrão (σb_{si})**	j	**Desvio Padrão (σb_{si})**
1	2200,84	2	810,043

Tabela 2 – Desvio padrão do modelo de nível de cinza

j	**Desvio Padrão (σb_{gi})**	j	**Desvio Padrão (σb_{gi})**
1	17018,1	2	13789,4
3	6848,04	4	6495,6
5	6231,78	6	5640,13
7	5076,28	8	4918,14
9	4637,89	10	4359,39
11	4041,06	12	3770,31
13	3678,94	14	3500,47
15	3436,83	16	3296,93

Tabela 3 – Desvio padrão do modelo de aparência

j	**Desvio Padrão (σc_i)**	j	**Desvio Padrão (σc_i)**
1	17019	2	13807,4
3	6854,21	4	6498,25
4	6498,25	4	6498,25
6	5669,53	6	5669,53
7	5093,71	8	4925,19
9	4641,34	10	4434,13
11	4049,26		

3.3 Resultados obtidos

Como imagem média do modelo, ou seja, a imagem que é possível gerar pela equação (13), considerando $c_i = c_0$, um vetor nulo, obtém-se a Figura 5:

Figura 5 – Imagem média do modelo

Aplicando a equação (5) em vetores b_{si} variando os dois principais componentes do modelo com ±3 vezes o desvio padrão da Tabela 1, pode-se obter o conjunto de imagens da Tabela 4, a seguir.

Tabela 4 – Imagens obtidas com a variação da forma

j	$-3\sigma b_{si}$	$+3\sigma b_{si}$	j	$-3\sigma b_{si}$	$+3\sigma b_{si}$
1			2		

Nota-se que o modelo tem poucos parâmetros de forma, o que é esperado, pois o modelo consiste de uma única pessoa com formato único de cabeça. Os parâmetros obtidos são de rotações.

Já utilizando a equação (9) em vetores b_{gi} variando os dezesseis principais componentes do modelo com ±3 vezes o desvio padrão listado na Tabela 2 obtém-se o conjunto de imagens da Tabela 5:

Tabela 5 – Imagens obtidas com a variação do nível de cinza

j	$-3\sigma b_{gi}$	$+3\sigma b_{gi}$	j	$-3\sigma b_{gi}$	$+3\sigma b_{gi}$
1			2		
3			4		
5			6		
7			8		

9			10		
11			12		
13			14		
15			16		

As variações de nível de cinza obtidas são principalmente combinações de movimentos dos olhos e lábios e sutis movimentos de expressões faciais. Também apresentam variações de luminosidade vindas das imagens originais.

E finalmente com a equação (13) aplicada em vetores c_i variando os onze principais componentes do modelo com ±3 vezes o desvio padrão citados na Tabela 3, tem-se o conjunto de imagens da Tabela 6:

Tabela 6 – Imagens obtidas com a variação da aparência

j	$-3\sigma c_i$	$+3\sigma c_i$	j	$-3\sigma c_i$	$+3\sigma c_i$
1			2		

3			4		
5			6		
7			8		
9			10		
11					

Já neste caso de variação da aparência se nota claramente a combinação dos modelos de forma e nível de cinza. Alterando unicamente os parâmetros de aparência, é possível obter ótima aproximação das variações dos outros modelos.

4 Discussões e conclusões

Neste artigo foi descrito um método para geração de poses de faces, e como exemplo foi utilizada a face do presidente americano Barack Obama. Para tanto foi feito uso da técnica de Active Appearance Model (AAM), que é bem conhecida como técnica de reconhecimento de objetos ou faces.

Foi obtido um modelo para o conjunto de dados por meio de um conjunto de matrizes de transformação. Alterando um conjunto de parâmetros internos e aplicando o modelo, podem-se gerar expressões faciais de determinada pessoa como se queira. Os resultados obtidos se assemelham com os obtidos na literatura, como visto na Figura 3, no entanto na literatura nota-se a alteração da pessoa e neste trabalho ocorre apenas a alteração das expressões.

É sugerido que a técnica obtida neste trabalho seja utilizada para geração de avatares para uso em ambientes virtuais como jogos ou redes sociais. Para futuros trabalhos, pode-se buscar a aplicação deste modelo com três coordenadas de cores, não apenas o nível de cinza, a fim de serem geradas imagens coloridas. É possível também a trabalhos futuros a aplicação de um modelo 3D como citado na literatura.

Referências

Abboud, B., Davoine, F. and Mo Dang. (2003) “Expressive Face Recognition and Synthesis”, Computer Vision and Pattern Recognition Workshop. CVPRW Conference, p. 54.

Abboud, B., Davoine, F. and Mo Dang. (2003) “Statistical modeling for facial expression analysis and synthesis”, Image Processing. ICIP. Proceedings of International Conference. Vol. 1, p. 653-6.

Cootes, T.F. and Taylor, C.J. (1994) “Modelling object appearance using the grey-level surface”, 5th British Machine Vision Conference, p. 479-88.

Cootes, T.F., Edwards, G.J. and Taylor, C.J. (1998) “Active Apperance Models”, ECCV ’98 Proceedings of the 5th European Conference on Computer Vision. Vol. 2, p. 484-98.

Cootes, T.F., Taylor, C.J., Cooper, D. H. and Graham, J. (1995) “Active Shape Models – Their Training and Application”, Computer Vision and Image Understanding. Vol. 61, p. 38-59.

Edwards, G.J., Taylor, C.J. and Cootes, T.F. (1998) “Interpreting face images using active appearance models”, Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference, p. 300-5.

Jianglong, C., Ying, Z. and Zengfu, W. (2007) “Facial Expression Analysis and Synthesis: a Bilinear Approach”, Information Acquisition. ICIA. International Conference, p. 457-64.

Lee, C.-S., Elgammal, A. and Metaxas, D. (2006) “Synthesis and Control of High Resolution Facial Expressions for Visual Interactions”, Multimedia and Expo, IEEE International Conference, p. 65-8.

Petajan, E. and Graf, H.P. (1996) “Robust face feature analysis for automatic speechreading and character animation”, Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference, p. 357-62.