Is that a Smile? Gaze Dependent Facial Expressions

Vidya Setlur   **Bruce Gooch**

Department of Computer Science
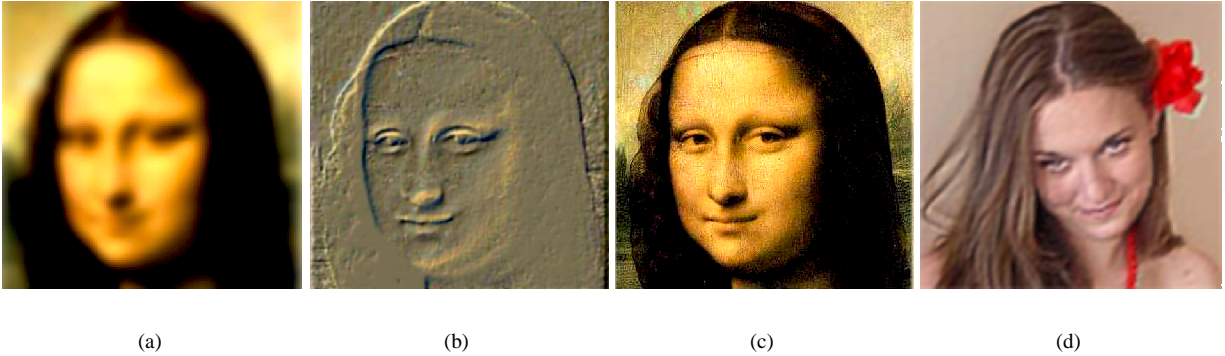Northwestern University

Figure 1: *a) A low spatial frequency filter reveals a more prominent smile. b) A more neutral expression is seen under high spatial frequency. c) The original image of Mona Lisa, courtesy http://www.allposters.com. d) An image generated by our system.*

Abstract

Based on artistic methods used for manipulating perception, we present a technique that creates facial images with conflicting emotional states at different spatial frequencies. The foveal and peripheral components of the human visual system tend to interpret emotional states differently, adding a degree of *elusiveness* to the facial image. Our technique first isolates the coarser low spatial frequency components and finer high spatial frequency details from two images with differing facial expressions. We then perform image segmentation with edge detection, and morph the images. In practice we have found that high spatial frequency elements determine the dominant expression in the resulting image, while the low spatial frequency elements contribute subtlety.

Keywords: spatial frequency, visual perception, facial expression, emotion, acuity, cognitive dissonance

1   Introduction

For centuries the famous Mona Lisa painting has perplexed art lovers with her enigmatic smile. Livingstone [2002] says that the smile becomes apparent when the viewer looks at other parts of the painting and disappears when the lips are looked at directly because of the way in which the human eye processes visual information. The center of gaze, called the "fovea," has a higher density of cones than anywhere else on the retina with highest visual acuity, capable of sensing fine detail. Peripheral vision on the other hand, is optimized for sensing coarser information such as, movement, and depth.

Figure 2: *Renoir's 'Luncheon of the Boating Party'. The painting's most amazing feature is its ability to reach its audience with the expression of the faces and the casual chaos of the party, while having no qualities of crisp lines or details. Image courtesy http://www.allposters.com.*

Livingstone explains the enigmatic nature of Mona Lisa's smile by the fact that "her smile is present almost entirely in low spatial frequencies, and is best seen by our peripheral vision". From Figure 1 it is apparent that there is a change in expression from low to high spatial frequency. Da Vinci used the shadows from her cheekbones to accentuate the Mona Lisa's mouth, making her smile more pronounced when viewed indirectly. This goes against the popular
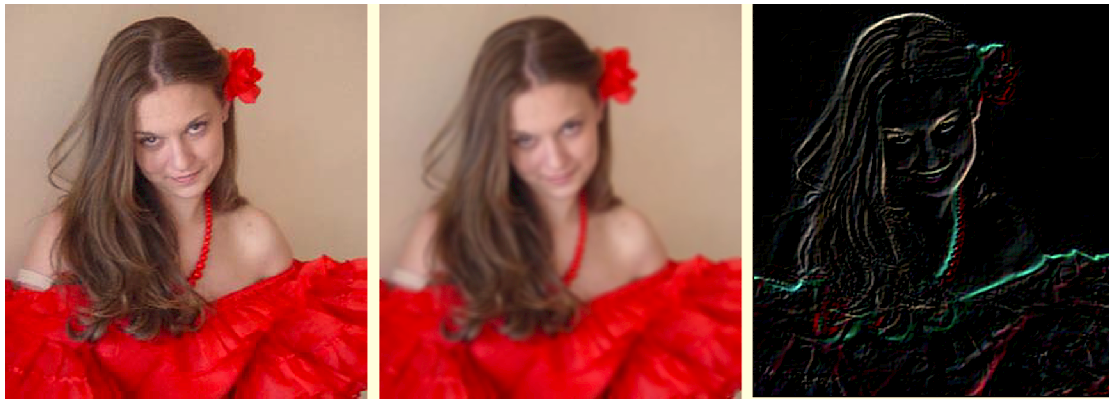
Figure 3: *From left to right: Applying spatial filtering. The original image. The image I_{low} after applying low spatial Gaussian filtering with $\sigma = 3$. The difference image of I_{high} after removing low frequency elements.*

belief that the mouth was blurred (a.k.a. sfumato) to make her expression ambiguous. Art historian Sir Ernst Gombrich [1995] has said, "Sometimes she seems to mock at us, and then we catch something like sadness in her smile. This sounds rather mysterious, and so it is; that is so often the effect of a great work of art."

Interestingly, Impressionist artists like Monet and Renoir also exploited the loss of precise spatial information in peripheral vision to introduce vitality into their art. The technique of *illusory conjunction*, a phenomenon where the color of one object is assigned to the shape of an adjacent object, creates a similar disconnect in the visual system causing it to complete the picture differently with each glance. Similarly, *equiluminance* occurs when colors occurring together have no brightness differences, but only hue differences. The edges of objects where the object and background have the same luminance, cannot be easily perceived. Equiluminant colors make a painting appear unstable as they seem to shift position, indicating motion. In the computer graphics community, Freeman and Adelson have developed a technique to produce motion without movement [Freeman et al. 1998]. Though the objects in their images do not actually move around, they manage to imply the impression of motion. This method can be used to convey motion without disturbing the spatial organization of the image.

The main motivation of our work is to allow non-artists to be able to easily create novel and interesting facial expressions. Psychological studies [Ekman 2003] have found cognitive dissonance occurs in the mind of the viewer as the human visual system attempts to rationalize between these different emotions, making these facial expressions more captivating. Goals of this research have been to implement the technique, and to allow the user to apply various spatial filters to each of the two source images before they are morphed. In order to generate such facial expressions, morphing tends to seamlessly integrate the low and high frequency components. Using Livingstone's theory as a basis, we consider this blend of low and high spatial elements as key for the generation of expressive faces. The user could choose the frequency and morphing weight parameters, based on the image resolution, and the facial expression desired. The final image is rendered based upon these parameters to generate a variety of artistic facial expressions.

2 Previous and Related Work

Creating meaningful and convincing expressions is one of the motivations behind the various approaches to the synthesis of facial expressions. One technique called *expressive expression mapping with ratio images* [Liu et al. 2001], maps illumination changes of one person's expression onto another person's image. These facial expressions exhibit subtle changes in illumination and appearance.

A broader set of methods target the generation of photorealistic human facial images. Algorithms have been developed to automatically construct functionally realistic models by considering the anatomical structure within a dynamic skin model of a human [Lee et al. 1995]. Photorealistic textured models have also been created by extracting texture maps from photographs of a human subject from multiple views [Pighin et al. 1998]. This approach however, involves manually marking initial corresponding points on the face from all of these different views. Another system [Guenter et al. 1998] captures three-dimensional geometry, color and shading information for human facial expressions. This is then replayed as a 3D talking head with a deformable face model complete with a changing texture map, producing lifelike reconstructions of facial expressions recorded from live actor's performances. A different technique stems from an example set of 3D face models to derive a morphable model by transforming the shape and texture of the examples into a vector space representation [Blanz and Vetter 1999]. Parameters such as gender, and fullness of face is used to perform face manipulations.

Other techniques produce facial expression animations for 3D models. One approach is to use existing animation data in the form of 3D motion vectors [Noh and Neumann 2001]. This technique allows animations created by any other tools to be re-targeted to new models. Another method, *expressive textures* [Fei 2001] uses textures of images of both synthetic faces or faces captured from video and of a simple low-polygon face model, concentrating on creating facial expression by manipulating the texture.

However, very little work has been done on facial expressions using human perception principles. Our motivation lies in the connections between human visual perception and the art of picture production [Durand et al. 2002]. Perceptual and cognitive principles have been used in various pictorial techniques by artists to create ambiguity in expression. In Caravaggio's *Bacchino malato*, the expression is strangely interesting. The young Caravaggio with a slight greenish complexion, is eating a bunch of small grapes, crouching behind the modest table where a couple of apricots try to make up for the void. The expression appears to display a degree of ambiguity with collective interpretations of anger, ill-health, or perhaps a slight smile.

We attempt to draw a common ground between perception and non-photorealistic rendering in order to develop an algorithm for the production of images that are not necessarily photorealistic, but are "effective", "expressive", or even "beautiful".

3 Implementation

We present a computational approach to artistic techniques used to manipulate the visual system. Our system first identifies the low and high spatial frequency features from each of the two source images (Section 4). Feature extraction is subsequently performed to form corresponding shape warps required for morphing. Finally, a unique facial expression is created by morphing the spatially varying features together with a color blending function (Section 5).

4 Spatial Frequency

Spatial frequency is the number of changes in brightness value per unit distance in any part of an image. Low spatial frequency is characterized by being tonally smooth with gradual changes. High spatial frequency on the other hand, is characterized by being tonally rough with abrupt changes. Most Impressionist works have soft lines, making it difficult to break the image into regions. Yet these paintings also contain areas of crispness that can accentuate a particular object in the painting. Figure 2 shows a well-known Impressionist painting that exhibits this quality. We use this technique as a basis for extracting low and high spatial frequency components from two source images. In this paper, we use the notations I_{low} and I_{high} for image sources of low and high frequencies respectively. These two spatial frequencies play an interesting role in producing a dominant facial expression with underlying subtlety.

Figure 4: *From top to bottom: The first two are the source images. The middle two images are generated after applying zero detection with LoG. The last two are generated after performing edge detection.*

4.1 Low-pass Filtering

To identify low spatial elements the original image is convolved with a Gaussian convolution kernel (of the pixel spreading type) to create a blurred image. The pixel spreading kernel is the conceptual equivalent of a fat paintbrush. For every new pixel it creates, it spreads out and solicits contributions from the surrounding pixels, while de-emphasizing the contribution of the center pixel. The Gaussian distribution in 2-D form is:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

where σ is the standard deviation of the distribution.

Convolving this Gaussian kernel with the source image I_{low}, we get M_{low}. This is expressed as:

$$M_{low} = (I_{low} * G),$$

where $*$ is the convolution operator.

4.2 High-pass Filtering

We perform high-pass filtering by finding the image difference between the source image I_{high} and I_{high} with a low-pass Gaussian filter G applied to it, i.e.,

$$M_{high} = I_{high} - (I_{high} * G)$$

Since M_{high} is the difference image, M_{low} and M_{high} are added together without altering the amplitude at any frequency. This also prevents color distortion that might occur if the high-frequency scaling is set too high.

Figure 3 shows images after these two spatial filtering techniques have been applied.

5 Feature Correspondences

We are particularly interested in determining the positional correspondences of various feature lines that outline the face, and body in the two source images. This standard geometry of image outlines is calculated using a zero-crossing feature detector with a Laplacian of Gaussian(LoG) operator. Edge detection is performed in the spatial domain, because it is computationally less expensive and often yields better results.

We first convolve the Gaussian smoothing filter with the Laplacian filter. This reduces the high frequency noise components prior to differentiation. We then convolve this hybrid filter with the image to achieve the required result. The 2D LoG function centered on zero and with Gaussian standard deviation σ has the form:

$$LoG(x,y) = -\frac{1}{\pi\sigma^4}[1 - \frac{x^2+y^2}{2\sigma^2}]e^{-\frac{x^2+y^2}{2\sigma^2}}$$

After the LoG filter is applied, the zero–crossing detector looks for places in the Laplacian of the image where the value of the Laplacian passes through zero, i.e. points where the Laplacian changes sign. Such points often occur at 'edges' in images, i.e. points where the intensity of the image changes rapidly. Zero crossings always lie on closed contours. Therefore the output from the zero–crossing detector is usually a binary image showing positions of the zero crossing points.

The threshold value used for the smoothing stage of this operator, strongly influences these zero crossings. In each of the four relevant directions, we look for a change in sign between the two opposite pixels on either side of the middle pixel. If there is a change the point is set to white and if there is no change it is set to black (Figure 4).

| (a) | (b) | (c) |

Figure 5: *Neutral-Slight Smile Example: Though the difference in expression between the first two images is very subtle, the result in the third image clearly shows a smile when looked at directly. a) An image with a neutral expression. b) An image with a slight smile. c) Combining the two by applying a 5X5 Gaussian mask with $\sigma = 2$ and morphing weight = 0.25.*

5.1 Line Detection

The system then performs line detection by applying the Hough Transform, to detect patterns of points on the binary images. The key idea is mapping a complex pattern detection problem into a simple peak detection problem in parameter space [Leavers 1992]. If we take a point (x', y') in the image, all lines which pass through that pixel have the form $y' = mx' + c$ for varying values of m and c.

The algorithm we use can be enumerated as follows:

1. Quantize (m, c) space into a 2D-array N for appropriate steps of m and c.

2. Initialize all elements of $N(m, c)$ to 0.

3. For each pixel (x', y') which lies on some edge in the image, we add 1 to all elements of $N(m, c)$ whose indices m and c satisfy $y' = mx' + c$.

4. Search for elements of $N(m, c)$ which have large values.

5. Check to see if the number of pixels on the line exceeds a threshold value. This is to eliminate negligible lines (e.g. tiny hair strands) that do not contribute significantly to the basic face outline.

6. Each one found corresponds to a line in the original image.

We use this method since Hough Transforms have the property that collinear pixels need not be contiguous. This proves to be very useful because quite often the system needs to detect lines with short breaks in them due to noise. We then consider the two-dimensional array N used for Hough transformation to map corresponding line segments on both the images. A neighborhood around each feature from the first image is mapped onto the second image. A similarity measure S is computed to measure how different the features in the second image is from the feature in the first.

$$S = \frac{1}{w_0(l_{low} - l_{high})^2 + w_1(\theta_{low} - \theta_{high})^2}$$

w_i are weights, and the length (l), orientation θ are different measures of the feature in the first(I_{low}) and second(I_{high}) images. The features with the highest similarity measure are considered to be the best match. Although this method uses very simple heuristics and could pose certain problems with accuracy, it tends to not affect our results drastically as the two source images considered for edge detection are of the **same** subject with **different** facial expressions, with their shape warps approximately the same.

5.2 Morphing

We use feature-based image metamorphosis [Beier and Neely 1992] to morph the two spatially filtered images. When used effectively, this technique can give the illusion that the images are transforming in a fluid, surrealistic, and often dramatic way. The feature line correspondences determine the image warp and the two images are warped to line-up as images I_A and I_B.

$$I_A = warp_A(I_{low})$$

$$I_B = warp_A(M_{low}) + warp_B(M_{high})$$

Once the images are warped we need to compute the blending of colors from the image combination. We implement linear cross-dissolving, which is of the form:

$$I = (1 - t)I_A + tI_B$$

where I is the intensity of the destination pixel, I_A is the intensity of source image A, I_B is the intensity of source image B, and t is the time in the morphing sequence, and $0 \leq t \leq 1$. The cross-dissolving transitions along the original first image I_{low}, and gradually adds in the high frequencies of the second, fading those out in the first. The morphing weight influences the ratio of high and low spatial features in the morphed image that ultimately determines which expression dominates over the other. Since the pixels in the source images are spatially filtered, the linear cross-dissolve sequence can introduce visual artifacts of highly contrasted groups of pixels, particularly near the edges. We therefore apply a technique of linear spatial filtering [Iwanowski and Serra 1999] on the morphed image. We applied a $3x3(s = 1)$ Gaussian mask for blending. In Figure 6, one can observe that as we increase the morphing weight, the facial expressions vary. Figure 5, is an example similar to the Mona Lisa, where we combine two expressions; one being neutral, and the other being a slight smile.

Figure 6: *From left to right: Gradual change in facial expression from sad to angry. The morphing weight starts at 0 at the first image and increases in increments of* 0.25 *till it reaches* 1.0 *at the last image, with a 5X5 Gaussian kernel and* $\sigma = 3$.

6 Results

One of the key results of our approach is that non-artists can implement artistic techniques to create facial expressions. We tested our software on several combinations of facial expression. For our case study, we used standard emotions that people normally perceive and understand [Ekman 2003]: *Sadness/agony, anger, surprise/fear, disgust/contempt, enjoyment.*

Cognitive theory suggests that the interpretation of facial expression is influenced by certain visual stimuli that are characteristic of a particular facial expression. In Figure 7 we can see that the angry expression is indicated by the knitting of the eyebrows and the tense mouth. The happy expression on the other hand is characterized by the raised eyebrows, lifting up of the cheeks and the smile. A combination of angry and happy expressions result in an interesting combination with the stretching of the subject's cheek for a smile, but a crease between the brows for anger as shown in Figure 7. A "haughty" expression is portrayed in the result in Figure 9 with a tense mouth, narrowed eyes and raised eyebrows. Our technique works well with a variety of facial expressions. However, it can produce artifacts in the image in cases where the subject's mouth is open or the teeth are visible for instance in Figure 8.

7 Conclusion

This work demonstrates that the spatial frequency information in an image of a face can be systematically changed in order to create gaze dependent expressions. We were able to show how spatial frequency and the morphing weight could influence the outcome of facial expression combinations. A dominant expression with an underlying subtlety seems to add an elusive quality to facial images.

Future directions for research include applying similar techniques to video. In addition, we believe that evaluating the results of our technique in psychophysical studies may prove interesting. For example, how do the images that result from this technique compare to standard photographs in terms of learning speed and accuracy.

8 Acknowledgements

References

BEIER, T., AND NEELY, S. 1992. Feature-based image metamorphosis. In *Proceedings of SIGGRAPH 1992*, ACM Press / ACM SIGGRAPH, Computer Graphics Proceedings, Annual Conference Series, ACM.

BLANZ, V., AND VETTER, T. 1999. A morphable model for the synthesis of 3d faces. In *Proceedings of SIGGRAPH 1999*, ACM Press / ACM SIGGRAPH, Computer Graphics Proceedings, Annual Conference Series, ACM.

DURAND, F., AGRAWALA, M., GOOCH, B., INTERRANTE, V., OSTROMOUKHOV, V., AND ZORIN, D. 2002. Perceptual and artistic principles for effective computer depiction. In *Course 13, SIGGRAPH 2002, San Antonio, Texas*, ACM.

EKMAN, P. 2003. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Henry Holt and Company, Incorporated.

FEI, K. 2001. Expressive textures. In *Proceedings of the 1st international conference on Computer graphics, virtual reality and visualisation*, ACM Press / ACM SIGGRAPH, ACM.

FREEMAN, W. T., ADELSON, E. H., AND HEEGER, D. J. 1998. Motion without movement. In *Proceedings of SIGGRAPH 1998*, ACM Press / ACM SIGGRAPH, Computer Graphics Proceedings, Annual Conference Series, ACM, 27–30.

GOMBRICH, E. H. 1995. *Story of Art*. Phaidon Press, Incorporated.

GUENTER, B., GRIMM, C., WOOD, D., MALWAR, H., AND PIGHIN, F. 1998. Making faces. In *Proceedings of SIGGRAPH 1998*, ACM Press / ACM SIGGRAPH, Computer Graphics Proceedings, Annual Conference Series, ACM.

IWANOWSKI, M., AND SERRA, J. 1999. Morphological interpolation and color images. In *Proceedings of the 10th International Conference on Image Analysis and Processing*, IEEE Computer Society, IEEE.

LEAVERS, V. F. 1992. *Shape Detection in Computer Vision Using the Hough Transform*. Springer-Verlag New York, Incorporated.

LEE, Y., TERZOPOULOS, D., AND WATERS, K. 1995. Realistic modeling for facial animation. In *Proceedings of SIGGRAPH 1995*, ACM Press / ACM SIGGRAPH, Computer Graphics Proceedings, Annual Conference Series, ACM.

LIU, Z., SHAN, Y., AND ZHANG, Z. 2001. Expressive expression mapping with ratio images. In *Proceedings of SIGGRAPH 2001*, ACM Press / ACM SIGGRAPH, Computer Graphics Proceedings, Annual Conference Series, ACM.

LIVINGSTONE, M. 2002. *Vision and Art: The Biology of Seeing*. Harry N. Abrams.

NOH, J., AND NEUMANN, U. 2001. Expression cloning. In *Proceedings of SIGGRAPH 2001*, ACM Press / ACM SIGGRAPH, Computer Graphics Proceedings, Annual Conference Series, ACM.

PIGHIN, F., HECKER, J., LISCHINSKI, D., SZELISKI, R., AND SALESIN, D. H. 1998. Synthesizing realistic facial expressions from photographs. In *Proceedings of SIGGRAPH 1998*, ACM Press / ACM SIGGRAPH, Computer Graphics Proceedings, Annual Conference Series, ACM.
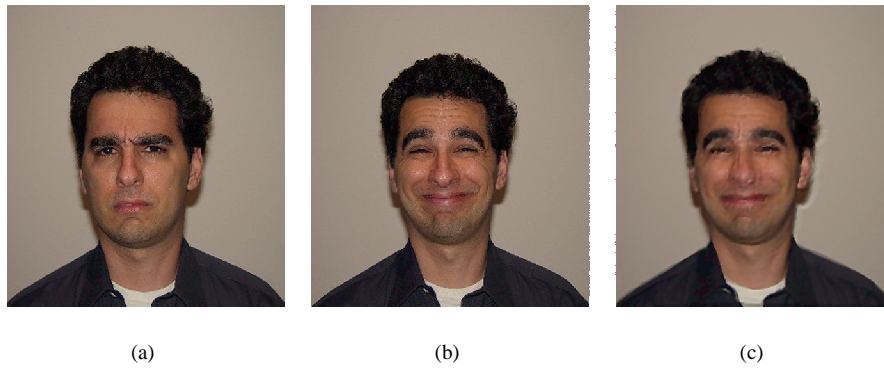
Figure 7: *An angry and happy expression morphed with a morphing weight of 0.75, using a 5X5 Gaussian mask with $\sigma = 3$. The result shows a mixture of both emotions, but dominated by the anger stimulus.*
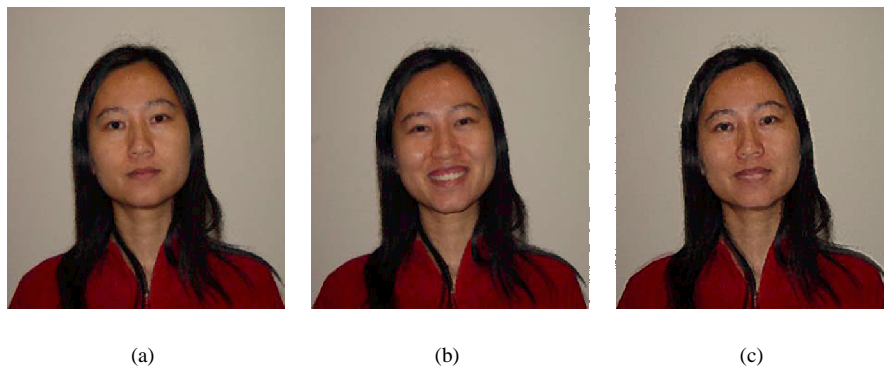

Figure 8: *The display of teeth in the source image leads to a ghost image in the final result.*

Figure 9: *Neutral and angry expressions are morphed at a morphing weight of 0.25, with a 3X3 Gaussian mask and $\sigma = 2$.*