

Visual Stimulus for Aural Pleasure

Gareth Davies, Stuart Cunningham & Vic Grout
Centre for Applied Internet Research (CAIR), University of Wales, NEWI
Plas Coch Campus, Mold Road, Wrexham, LL11 2AW, North Wales, UK
mrgdavies@googlegmail.com | s.cunningham@newi.ac.uk | v.grout@newi.ac.uk

Abstract. Can *images* be represented by *music* and if so how? This is one of the principal questions addressed in this paper through the development, implementation and analysis of computer generated music and assessment of the music produced.

Images contain a multitude of *definable information content* such as the colour values and dimensions, which can be analysed in various statistical combinations to produce histograms and other interpretations of the content. This provides music composition algorithms with a variety of data sets and values which can be used to inform parameters used in the music generation functions. However, images also contain *semantic information* which is widely open to interpretation by a person looking at the image. Images, much like music, are able to induce *emotions, feelings* and other *soft responses* in a viewer, demonstrating that images too are more than simply the sum of their parts. Therefore, if we are to generate music from image data then a goal of doing so would be to produce music which can begin to invoke similar *humanistic* responses.

The work presented in this paper demonstrates how compositional algorithms can be used to create computer generated music from a set of various images. Several established music composition algorithms are explored as well as new ones, currently being developed by the authors of this paper. We explore different uses and interpretation of the image data as well as different genres of music, which can be produced from the same image, and examine the effect this has. We provide the results of listener tests of these different music generation techniques that provide insight into which algorithms are most successful and which images produce music that is considered to be more popular with listeners.

Finally, we discuss our future work, directions and the application areas where we feel that our research would bring particular benefit. In particular, we seek to incorporate the work presented in this paper with other technologies commonly used to assess visual and psychological responses to images and propose techniques by which this could be harnessed to provide a much more dynamic and accurate musical interpretation of an image. One of the main goals we aim to achieve through further development of this work is to be able to successfully interpret an image into a piece of music or sound, which could be played to a visually impaired or blind listener to allow them to grasp the emotional content and responses which imagery can invoke.

1. Introduction

The reflection and association between imagery and music is a common manifestation in the 21st Century, particularly due to the nature of the rich, multimedia environments and data to which we are exposed everyday. A prime example is the success of the music video and MTV, which has received worldwide recognition and familiarity in a relatively short period of time. The way in which music and images or video are often combined demonstrates the value and semantic link which humans can easily make and the extra information provided when these otherwise independent media are fused [1]. The emotional influence carried by this combination of these media and the imagery and interpretation associated are often thought to be highly powerful and influential in many areas, when interpreted by users into the ‘real world’ [2, 3, 4].

In this work, we examine the viability and practical success of analysing image data and creating musical compositions influenced by the image data.

The initial outcome of this research is to produce a functioning program that takes a bitmap image as the source file, reads the data and uses that data as the inputs to an algorithm to generate MIDI (Musical Instrument Digital Interface) music files. Users will have a choice of algorithms and musical genre. The choice of genre will affect the tempo and instrumentation in the MIDI file and the image data affects the scales, pitch, time signature, tempo and structure applied. In the longer term, one of our aims is to produce music which is reflective of the semantic content also present in the image. Such techniques could be used, for example, to create representations of images for those who are not fully able to interpret the image due to visual impairments.

We begin by discussing the information and data that is contained within images and how these factors can be employed when attempting to create music. In particular, we draw attention to the fact that images contain both numerical, raw data as well as a harder to define, emotional or humanistic content. We then discuss how image data can be interpreted to produce music and present several compositional algorithms which can be used to automatically generate musical sequences, including an original algorithm developed by the authors of this paper. In the subsequent section, these algorithms are tested with human listeners and the results discussed. Finally, we conclude the work completed to date and discuss areas which are currently under investigation and development to further improve and refine the methods and techniques detailed within this paper.

2. Images and Information

Humans gather a large majority of their information from sight. *A picture is worth a thousand words*¹. Images allow not only capture and representation of a moment or scene, but also provide a vehicle by which humanistic emotional responses can be triggered. Therefore, we consider images to hold two very distinct values when we interpret visual data and emotion into musical audio data and emotion.

¹ http://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words

2.1. Data Content

There are varying types of data that can be extracted from an image. For example, the RGB (Red, Green, Blue), CMY (Cyan, Magenta, Yellow) or HSL (Hue, Saturation, Luminance) values for each individual pixel, histograms, the width and height of the image, and other more meta properties of the properties of the file itself, the name, size on the disk, etc. The content of the image which describes the pixels and colour values which constitute the image are of particular use and these can be analysed through a number of statistical techniques to gain further depth and insight into the particular qualities of the image. Of such techniques, histograms in particular, provide a useful tool which allows us to easily convey large amounts of information about the colour content of an image.

A histogram provides a graphical representation of the frequency of a set of results. In the case of an image it would be the frequency of the occurrence of a particular colour. This allows us to obtain a distribution of the components which form the image. A histogram can be produced from individual values such as all the red values from an image or be made up of a combination of values such as the total RGB values or a combination of the CMY values. This yields a large amount of data from an image and the different permutations of the histogram can provide vastly different values. As an example, consider the greyscale image in Figure 1 and the associated histogram presented in Figure 2.



Figure 1 - Sample Greyscale Image

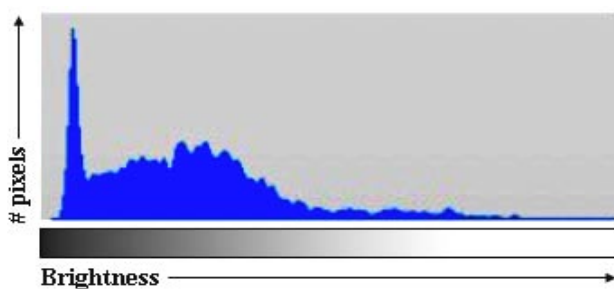


Figure 2 - Histogram of Greyscale Image

Such information about an image gives us initial tools with which to be able to reflect the content of the image in the music we produce. For example, it would be expected that particularly bright images and images with large variations in colour would translate into more diverse and intricate music.

2.2. Semantic Content

When we approach image analysis we must also consider the human and emotional responses which viewers will have to an image. We know that humans frequently exhibit fundamental emotional responses when viewing differing visual imagery [5]. It is worth thinking about the attraction and fascination that humans have with classical works of art such as Leonardo da Vinci's *Mona Lisa*. Consider also imagery which invokes many types of emotional response such as happiness, sadness, distress and contentment. Images provoke many kinds of emotional and humanistic responses and we must attempt to reflect these responses in the associated music produced from an image.

Such responses are much harder to quantify and predict using purely statistical analysis. For example, consider the expected emotional response from a viewer who observes an image of war, which might depict a bloody battle between two soldiers. In this case we would expect a viewer to have a negative emotional response such as sadness or horror. Now consider an image which might provoke a more positive emotional response, such as the image of a man giving a woman a bunch of roses in a grassy meadow. We would expect this to have an emotional response of happiness or compassion. However, if we consider the data present in these images we find many common features. For example, strong reds (blood in the battle image, the roses in the romantic image), flesh tones (the soldiers in the battle, the man and woman in the meadow), light blues (the sky in both images), grassy or earthy backgrounds (the ground in the battle, the meadow in the romantic image). We can surmise therefore, that the histograms extracted from these images would be very similar. This is not an ideal scenario and indicates that we need to consider techniques other than purely the raw data content in order to fully extract emotional content for music from images. We discuss methods of addressing this issue in our 'Conclusions and Future Work' section.

3. Creating Music from Images

Music has been composed algorithmically in the past and it can be traced back as far as the ancient Greeks as Grout and Claude stated:

"The word music had a much wider meaning to the Greeks than it has to us. In the teachings of Pythagoras and his followers, music was inseparable from numbers, which were thought to be the key to the whole spiritual and physical universe. So the system of musical sounds and rhythms, being ordered by numbers exemplified the harmony of the cosmos and corresponded to it" [6].

This notion from the Pythagoreans is not necessarily as abstract as it may sound. Autistic people that can do huge numerical calculations without the aid of calculators or computers often see numbers as colours and/or shapes. The fact that they can successfully carry out these calculations quickly indicates that there are associations between these visual objects and the numbers.

More recently, Mozart used a simple form of algorithmic composition in his *Musikalisches Würfelspiel* or *Dice Music*, where he had written a number of musical parts that he put together with the aid of dice rolls into a complete musical piece. Algorithmic composition with computers began in the 1950's and 1960's. Notably a piece called the *Illiac Suite* (1957) was produced in the University of Illinois, USA [7]. These approaches demonstrate how simple tools such as random variables, probabilities, transition tables, and rules can be employed in the generation of musical elements and the structuring of therein. We chose to employ three compositional algorithms in conjunction with the image data.

3.1. Image Selection

The images selected for our investigations play an important part when assessing the results and success of our study. To provide a balanced outlook on the production of music from images, a set of reference images was chosen to be employed in all tests. These sets of images formed four categories: people, city, landscape, and art. This spread allows us to accommodate a large variety of images common in everyday life. This selection should provide the compositional process with differing ranges and distributions of colour values. Samples of images from each of the categories can be seen in Figures 3 to 6.



Figure 3 - People Category



Figure 4 - City Category

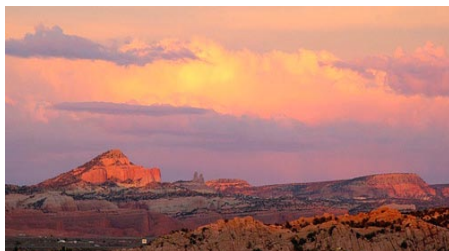


Figure 5 - Landscape Category



Figure 6 - Art Category

3.2. Assignment of Musical Properties

A primary issue which had to be dealt with, was to determine the best method to extract properties from the image which can be mapped against the parameters required for a musical composition. To determine which image colour values were best for selecting which MIDI data, a program was produced that generated the charts for the RGB and HSL values averaged over the height. This enables the flow of the song to be analysed given which colour values are used for which musical attribute. The various mapping we experimented with are detailed in Table 1.

| Variation | Tempo | Time Sig. Numerator | Time Sig. Denominator |
|-----------|------------|---------------------|-----------------------|
| 1 | Red | Green | Blue |
| 2 | Green | Blue | Red |
| 3 | Blue | Red | Green |
| 4 | Hue | Saturation | Luminance |
| 5 | Saturation | Luminance | Hue |
| 6 | Luminance | Hue | Saturation |

| Variation | Pitch | Duration | Velocity |
|-----------|------------|------------|------------|
| 1 | Hue | Saturation | Luminance |
| 2 | Saturation | Luminance | Hue |
| 3 | Luminance | Hue | Saturation |
| 4 | Red | Green | Blue |
| 5 | Green | Blue | Red |
| 6 | Blue | Red | Green |

Table 1 - Assignment of Image Attributes to Musical Properties

To complement this, a short excerpt from a song for each algorithm, using the various colour values for the various MIDI data, was produced and a questionnaire to accompany these excerpts was made. This questionnaire was issued to volunteer listeners who listened to the excerpts of the songs and then selected which one they preferred for each algorithm. The results of this test will influence the decision of how to map the image attributes to musical properties.

3.3. Compositional Algorithms

Once various interpretations of image data into musical properties were determined, these could then be used to inform the compositional algorithms. We created small libraries of fixed musical selections, such as various standard musical scales, time signatures, tempos, note duration, and velocity. It was necessary to apply these constraints to the ways in which the image information was interpreted and quantised into a form which was suitable for the algorithms, especially given the relative scale of the investigation. This discussion is detailed and goes beyond the scope of this paper. However, it should be remembered that such constraints were required in order to be able to successfully interpret values from images into musical properties. Roughly speaking the data obtained from images would have a broader numerical range than the musical parameters. Consider the range of 0 to 255 from an image attribute and time signature numerator which is usually much smaller, for example. The algorithms we used as part of our work are as follows.

3.3.1. Algorithm 1 – 1/f

Due to the self-similar nature of this algorithm and the fact that it has been frequently used previously for musical composition, the 1/f algorithm proved a logical selection to combine with image data. The 1/f algorithm has no direct input but has a random number generator which can be seeded with an integer, so this provides an indirect method for the algorithm to be affected by image data.

3.3.2. Algorithm 2 – Euclidian

The Euclidean algorithm takes two number values and returns the greatest common divisor, this accommodates the input and output that is required and was another obvious selection when we consider the amount of image data available. This can be simply demonstrated with the following sample of C code:

```
int GetEuclidian(int x, int y)
{
    int z;
    while(y!=0)
    {
        z = y;
        y = x%y;
        x = z;
    }
    return x;
}
```

3.3.3. Algorithm 3 – Colour Selection Process

Strictly speaking this is not an algorithm but more of a selection process. It was developed to provide an alternative to the two algorithms already selected.

The process works by using the actual colour values from the image. The theory behind this is that this should provide the most direct link between the image and the musical composition produced. Which colour values are used with which MIDI data will have an effect on the type of composition that is produced.

4. Discussion of Results

4.1. Assignment of Musical Properties

The results of the investigation into how to map image attributes to musical properties is show in Table 2. This shows the percentages of user preference for each colour/music assignment variation, tested against each of the three compositional algorithms.

| Variation | 1/f | Euclidian | Colour Selection | Average |
|-----------|--------|-----------|------------------|---------|
| 1 | 8.33% | 6.25% | 12.50% | 9.0% |
| 2 | 14.58% | 14.58% | 14.58% | 14.6% |
| 3 | 6.25% | 12.50% | 10.42% | 9.7% |
| 4 | 29.17% | 20.83% | 25.00% | 25.0% |
| 5 | 22.92% | 27.08% | 20.83% | 23.6% |
| 6 | 18.75% | 18.75% | 16.67% | 18.1% |

Table 2 - Result of Image Interpretation Variations

The configurations with the greatest number of responses for each of the three algorithms were chosen as follows:

- 1/f Algorithm – Variation 4
- Euclidian Algorithm – Variation 5
- Colour Selection Algorithm – Variation 4

An interesting point to note is that the majority of responses were selecting the HSL values to represent tempo and time

signature and the RGB values to represent pitch, duration and velocity; variation 4, which had the largest overall average of selection across all the tests. Also of note is that variations 2 and 6 produce almost equal results across all of the algorithms.

4.2. Compositional Algorithms

The algorithms currently available give the user a chance to evaluate how a different algorithm with the same image can make a completely different song and can enable them to make a choice about which algorithm makes the best representation of the image with sound. To enhance the ability to see each algorithms performance across the six variations, a series for each algorithm is plotted in Figure 7.

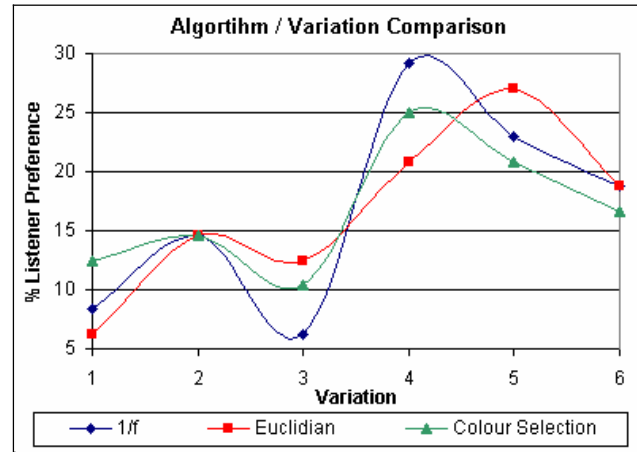


Figure 7 - Compositional Algorithms & Variations

At this stage, we are more focussed upon establishing suitable mappings between the image attributes and the musical properties and the trends of each series in Figure 7 suggest that our assumption and mapping proposed in the previous subsection is valid.

However, we are also interested in starting to investigate the effect that each of the algorithms may have had upon the user preferences. As we can see in Figure 7 the user selections for some algorithms take the form of a more varied distribution across each of the mapping variations, whilst others are less varied. The data presented in Table 3 also helps to reveal the properties of each of the algorithms across the six variations in terms of the percentage of user preferences.

| Algorithm | Min | Max | Variance | Std Dev |
|------------------|-------|-------|----------|---------|
| 1/f | 6.25 | 29.17 | 76.39 | 8.74 |
| Euclidian | 6.25 | 27.08 | 52.08 | 7.22 |
| Colour Selection | 10.42 | 25.00 | 29.51 | 5.43 |

Table 3 - Data Ranges in Results

If we consider the 1/f algorithm first, it has the most extreme fluctuations across the six variations. This would suggest that the 1/f algorithm is more influenced by the mapping of image attributes and therefore the types of image which are used to generate music using the algorithm. Interestingly, the 1/f method also contains the maximum value and shares the minimum with the Euclidian, in the data sample. Although further testing is required, this further enforces the belief that this technique is sensitive to the image mapping and attributes, but it also indicates that it could be considered to hold the most potential in producing satisfactory musical compositions.

The performance of the Euclidian algorithm tends to vary considerably across the six variations, although for variation 4 and 5 it exhibits almost inverse performance to that of the other two composition methods. However, in terms of the overall variation across all six variations, the results of the Euclidian method are almost as varied as that of the $1/f$ algorithm. This, and the fact that the Euclidian algorithm shares the lowest value in the sample data with the $1/f$ algorithm, suggests a similar dependency upon careful image attribute mapping and selection. The trend exhibited by our own Colour Selection process generally follows the same overall trend of the $1/f$ technique although it is not scaled to the same extremes. Our method seems to provide the most consistent, flattest results across each of the variations. The Colour Selection technique displays the smallest mathematical variation and standard deviation of all three of the algorithms. Therefore, whilst we consider that the $1/f$ algorithm and the Euclidian algorithm are very dependant on image attributes and data, our technique is the most likely to produce satisfactory musical compositions. However, these results suggest that we must consider improvements to this method if we intend to be able to produce excellent musical compositions.

5. Conclusions & Future Work

We recognise the immediate need to provide more detailed assessment of the three algorithms used and the image selections. In this work we were primarily concerned with determining the most suitable mappings of parameters and attributes. However, with the data and resources now available we can launch a more detailed investigation into determining the usefulness or listener ranking for each of the algorithms in terms of the music which is generated. The same also applies to exploring which of the image categories produces the most satisfying music. As in any study, where human perception and interpretation is required, it will always be challenging to ascertain such information as absolute fact, due to the varied nature of perception, and particularly musical preference, across a broad field of listeners. The key to producing substantiated evidence of these facts lies in employing a large range of tests and subjects to ensure as broad a listener base as possible, and to ensure that qualitative data is also produced at these times.

One of the key aspects of this investigation was the representation of images as music. This was achieved from a technical point of view; the software and algorithms will generate music from images and will have a different output for each image that is used. This has been proven in the form of the listening tests performed. It is also true that by using the different image data available for different aspects of the musical composition hugely different musical compositions were produced and the same also applies when using the same image with different compositional algorithms.

An interesting point which leads us to future investigation and development is how well the images are represented by the music produced. Also, will it ever be possible to *truly* represent images with music? Again this is affected by the way that people view images. Different people have vastly different opinions on how an image should be represented by music, for example people who are affected by the condition Synesthesia can hear music while they are looking at images, and these people would therefore likely have a strong opinion on what music would be best represent an image [8]. Other people to consider are those affected by colour blindness they are likely to have a completely different opinion to those who are not.

This leads us into a significant area to further our work to date; to monitor, and subsequently model, the ways in which human viewers look at an image. The ability to use eye-tracking

technology provides the research field with the opportunity to further investigate and correlate vision with emotive response [9]. We propose that by employing eye-tracking tools, it will be possible to further determine the most crucial and important parts of an image when viewed by a fully-sighted viewer. Our hypothesis is that the emotional response experienced by the viewer of an image must be directly related to the particular parts of the image which the viewer focuses on as these are the principal stimuli which invoke the emotional response. Work by Jackson *et al.* substantiates this hypothesis [5]. Although their work is more concerned with psychopathology and psychophysiology their work demonstrates distinct correlation between emotional responses to images and the reflection of emotional response in the eyes. This may also help to address the example cited in section '2.2 Semantic Content' where we discuss the issue related to purely statistical analysis of image data in the case of the two contrasting pictures; the scene of a bloody battle and the scene of a man and woman with a rose.

The notion of investigating specific focal points within images to determine the most important areas to a viewer is reinforced by the works of Santella and DeCarlo [10, 11]. Their work is particularly focussed on obtaining and attempting to model the human perception of images, such that which is available in photographs. Although the goal of Santella and DeCarlo's work is to modify and create more abstract interpretations of the images analysed, they focus on being able to ensure that the most significant portions and areas within the image receive the most attention and level of granularity within the new version. This reflects the intentions we propose; to focus music composition algorithms on the most significant areas within an image which hold the greatest value and importance to the viewer.

For example, consider the image shown in Figure 8.



Figure 8 - Image with Eye-Tracking Overlay

This image is taken from data repository of Santella and DeCarlo's work and the eye-tracking data associated with the image has been overlaid [10]. The circles show the areas upon which the viewer has focused. The size of the circle indicates the length of time the viewer spent observing that particular area, the larger the circle, the more time spent viewing that area. This could be considered a weight metric within the image.

We can see that in the case of this example, the viewer focuses primarily around the area of the flower and the butterfly, and only pays small attention to the background setting. We propose that as part of our future work it would be possible to record multiple viewing of images from fully-sighted users and determine the common areas of focus within images and the correlated emotional response. Once a sufficient perception

model has been created for an image these additional parameters could act as filters to influence the compositional algorithm. As a simple example, the eye-tracking data could be used to direct the composition techniques at specific parts of the image, rather than analysing the whole image. Through iterative testing and development it is hoped that this process would be refined so as to be able to produce the same emotional experience through a piece of music for a visually-impaired listener as a sighted viewer would have when they see an image. This moves us beyond the question “*What does the Mona Lisa sound like?*” towards attempting to address the question “*What does the Mona Lisa feel like?*”

Of particular interest when attempting to discern and assess suitable emotive responses to images would be to create music using images of *Rorschach inkblots*. The investigation in this case would be to see if the music produced by inkblots also invoked the same emotional functions and responses in listeners as the actual inkblot images produced when viewed. This would provide tangible evidence that the music being created from our compositional methods was indeed resulting in the expected human responses.

Another interesting possibility to allow further tailoring of the music produced to each individual user could be in the use of genetic algorithms. The user could provide the fitness evaluation needed to determine best fit, this way the music composition algorithms could learn an individual’s preference and attempt to match their representation of the image to that. However, obtaining suitable fitness metrics for visually impaired users could be unrealistic and therefore a more practical solution would be to base this metric on a suitable distillation, such as the mean, of a more easily obtainable data set, as we suggest previously with the eye-tacking data.

Additional intentions for future testing include integration and evaluation of more compositional algorithms and the integration of an interactive painting window, which would allow playback of images being drawn on screen by the user. This could be further enhanced by experimenting with real-time music generation whilst the user is drawing. This presents another intriguing area for future consideration.

References

- [1] Chen, T., Rao, R.R., *Audio-visual integration in multimodal communication*, Proceedings of the IEEE, Vol. 86-5, pp. 837 - 852 (1998).
- [2] Royet, J.P., Zald, D., Versace, R., Costes, N., Lavenne, F., Koenig, O., Gervais, R., *Emotional Responses to Pleasant and Unpleasant Olfactory, Visual, and Auditory Stimuli: a Positron Emission Tomography Study*, Journal of Neuroscience, Vol. 20-20, pp. 7752-7759 (2000).
- [3] Sherman, B.L., Dominick, J.R., *Violence and Sex in Music Videos: TV and Rock n' Roll*, Journal of Communication, Vol. 36-1, pp. 79 - 93 (1986).
- [4] Hansen, C.H., Hansen, R.D., *How rock music videos can change what is seen when boy meets girl: Priming stereotypic appraisal of social interactions*, Sex Roles, Vol. 19-5/6, pp. 287 - 316 (1988).
- [5] Jackson, D.C., Malmstadt, J.R., Larson, C.L., Davidson, R.J., *Suppression and enhancement of emotional responses to unpleasant pictures*, Psychophysiology, Vol. 37-04, pp. 515 - 522 (2000).
- [6] Grout, D.J., Claude, V.P., *A History of Western Music*, 5th Edition, New York, W. W. Norton & Company (1996).
- [7] Maurer IV, J.A., *The History of Algorithmic Composition*, Stanford University, USA (1999). Available at: <http://ccrma-www.stanford.edu/~blackrse/algorithm.html> [Accessed: 20th August 2007].
- [8] Harrison, J., *Synaesthesia: The Strangest Thing*, Oxford University Press, Oxford, UK (2001).
- [9] Kiegler, K., Moffat, D.C., *Investigating the effects of music on emotions in games*, Proceedings of Audio Mostly Conference, Piteå, Sweden (2006).
- [10] Santella, A., DeCarlo, D., *Abstracted Painterly Renderings Using Eye-Tracking Data*, Proceedings of the 2nd international symposium on Non-Photorealistic Animation and Rendering (NPAR), ACM Press, New York, USA (2002).
Data set available at: <http://www.cs.rutgers.edu/~decarlo/data/npar02/index.html> [Accessed: 21st August 2007]
- [11] DeCarlo, D., Santella, A., *Stylization and Abstraction of Photographs*, Proceedings of ACM SIGGRAPH 2002, New York, USA (2002).