Kovesi, P. (2009). Video Surveillance: Legally Blind?
In Digital Image Computing: Techniques and Applications, 2009. DICTA '09. pp. 204–211.
IEEE Publishing. https://doi.org/10.1109/DICTA.2009.41

Video Surveillance: Legally Blind?

Peter Kovesi
School of Computer Science & Software Engineering
The University of Western Australia
Crawley, Australia
Email: pk@csse.uwa.edu.au

***Abstract*—This paper shows that most surveillance cameras fall well short of providing sufficient image quality, in both spatial resolution and colour reproduction, for the reliable identification of faces. In addition, the low resolution of surveillance images means that when compression is applied the MPEG/JPEG DCT block size can be such that the spatial frequencies most important for face recognition are corrupted. Making things even worse, the compression process heavily quantizes colour information disrupting the use of pigmentation information to recognize faces. Indeed, the term 'security camera' is probably misplaced. Many surveillance cameras are legally blind, or nearly so.**

***Keywords*-CCTV; surveillance; face recognition; image quality; spatial resolution; compression;**

I. Introduction

This paper arises from many frustrating years attempting to assist police with the enhancement of numerous surveillance video images. Almost invariably I was unable to to do anything useful. While it is fairly straightforward to make a poor quality video image look nicer, it is exceedingly difficult to enhance an image to the point that you are able to see something that you could not see before. Occasionally some small successes were achieved [16].

It was somewhat belatedly that I realized that the quality of surveillance video was much worse than I had realized, and any attempts at enhancement probably futile. This point was made clear to me when I was approached to enhance the image shown in Figure 1. In this example the image was as good a quality as one could ever expect from a surveillance camera, so there was not much one could do. It was well illuminated and free from noise, but any number of people could have been matched with the image! Nevertheless with a small amount of contrast stretching, a bilinearly interpolated enlargement of the image was sent to court. Despite my misgivings this poor quality image proved its worth because when the defendant was shown the images he immediately pleaded guilty! While this may be amusing it is not that surprising. We are very good at recognising even very poor quality images of ourselves. An image of yourself at the scene of a crime will also bring forward the emotions you were feeling at the time and it would be difficult to suppress and disguise these. If the defendant had been able to suppress these emotions and had said "That's not me" the prosecution would have had trouble arguing its case.

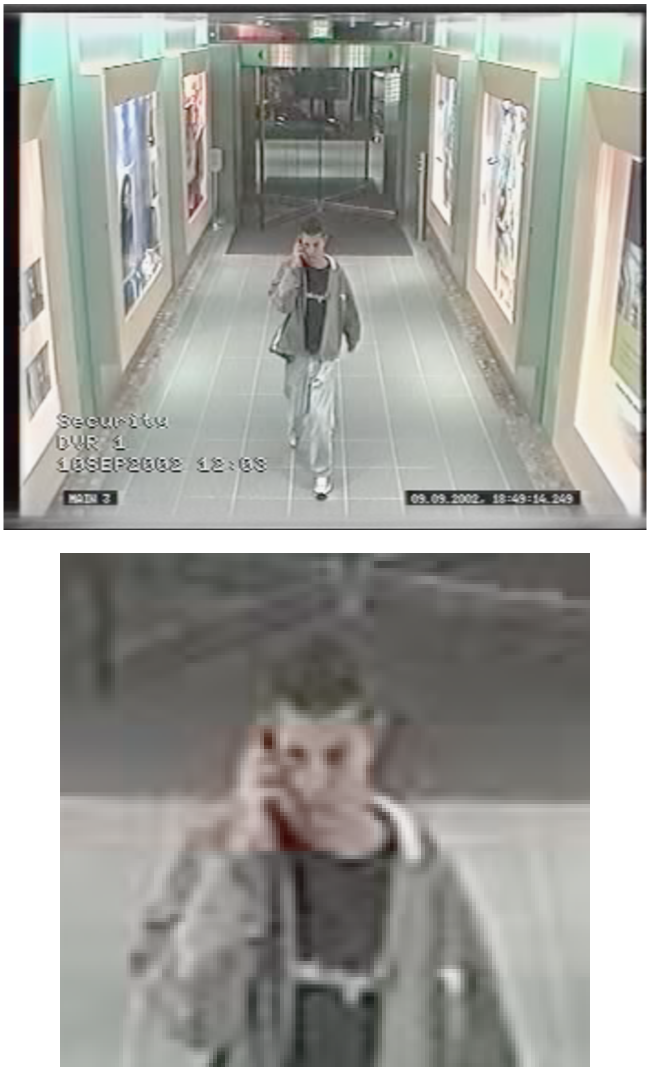

Figure 1. A good quality surveillance image and its bi-linearly interpolated enlargement — of any number of possible people.

This example then raised a number of questions.

- What image quality do we need for identification?
- How do you measure image quality?
- What is the image quality from a surveillance camera?

- What is the effect on image quality when you:
  - Record to video tape?
  - Use image compression?

Humans are actually rather poor at recognizing faces. This may seem counter-intuitive given that so much of our social life revolves around recognizing people and interacting with them on the basis of who we have recognized them to be. We are only good at recognizing faces that we are familiar with, such as our own, or of family and friends, and well known celebrities [20], [5]. The more familiar we are with a face the better our ability to recognise it in even a very poor image. However, when it comes to recognizing faces that we are not familiar with our performance is disturbingly poor. This has been demonstrated in a number of studies. Kemp, Towell and Pike [15] tested the value of having photos on credit cards. They found that when a user presented a card with a photograph of someone else that had some resemblance to the user, they were challenged less than 40% of the time. Bruce et al. [3], [4] tested the ability of people to match good quality CCTV images of unfamiliar faces under a variety of scenarios. They found correct recognition rates are typically only 70-80%. An illustration of one of their experiments is shown in Figure 2. Here they tested the ability of observers to match a good quality photo of a target person against 10 good quality CCTV images of faces. If the target was present the observer would have to correctly match the person, alternatively the observer would have to declare that no CCTV image of the subject was present. When the target was present in the array, 12% picked the wrong person and 18% said they were not present (overall only 70% correct). When the target was not present in the array 70% still matched the target to someone in the array.

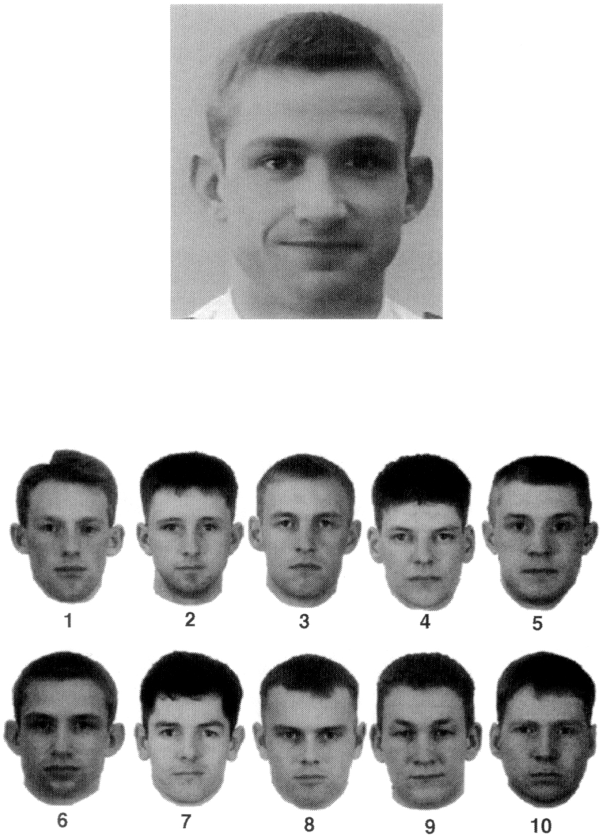

Figure 2. Example of test conducted by Bruce et al. [3]. Is this person in the array? If so, match the correct image.

If human face recognition is so poor, what is the state of automated face recognition? In the Face Recognition Vendor Test 2002 [10] the one-to-one verification results of the better performing systems produced a FRR of 0.2 at a FAR of 0.001. The 2002 one-to-many identification performance results for the better systems were roughly ∼90% for a database of about 100 individuals. This fell to ∼65-75% for a database of about 37,000 individuals. In this case the images were US visa application photos taken with standardized equipment and with white backgrounds, an ideal a situation as you could ever expect.

The Face Recognition Vendor Test 2006 results were an order of magnitude improvement [11]. The one-to-one verification results of the better performing systems, of faces under controlled lighting, produced a FRR of around 0.01 at a FAR of 0.001. These performances were achieved with high resolution images with the distance between the eyes being typically 350 pixels. The performance of some systems held up very well even on lower resolution images with around 75 pixels between the eyes. However, for faces under uncontrolled lighting and with high resolution images, these performances would fall to a FRR of around 0.15 to 0.3. So despite these very impressive achievements there is still some way to go, especially under uncontrolled lighting conditions.

The quality of images from surveillance cameras is typically well below that used in the Vendor Tests. A 768×576 PAL image is only 0.4 megapixels. This image is then recorded to video tape or heavily compressed and stored digitally. So then, what image quality is needed for face identification? Image quality is defined by many attributes, some of which might include

- Minimum feature size that can be resolved
- Noise level
- Quality of luminance reproduction
- Quality of colour reproduction.

This paper will mainly consider the spatial resolution requirements for face recognition and discuss some aspects of luminance and colour reproduction.

II. Spatial Resolution

A number of studies have looked at the spatial frequencies that are important for face recognition. Hayes, Morrone and Burr [14] concluded that the spatial frequencies most important for face recognition are around 20 cycles per face width. Costen, Parker and Craw [6] suggested that the range of 8 to 16 cycles/face width was the most important. More recently Näsänen [17] concluded that maximum sensitivity is centred around 8 to 13 cycles/face width and the bandwidth of the important spatial frequency range is just under two octaves.

While each study comes to slightly different conclusions the general outcome is that, for humans, face recognition is tuned to a set of spatial frequencies ranging from about 20 cycles per face width down to about 5 cycles per face width. Frequencies higher than this do not significantly improve recognition performance because they presumably only reveal insignificant features such as minor skin blemishes. At the other end of the scale, frequencies lower than 5 cycles per face width are perhaps mostly a function of lighting variations rather than facial features. Obviously to be able to recognize unfamiliar faces with some confidence one needs to be able to resolve spatial frequencies towards the upper end of this range. That is, frequencies greater than about 10 cycles per face width and preferably up to 20 cycles per face width

Given an average face width of about 160mm these two frequencies of 10 and 20 cycles/face width correspond to spatial wavelengths of 16mm and 8mm respectively. These correspond very nicely with the bar groupings at the top-right and bottom-right of the USAF chart, as shown in Figure 4. The 1951 USAF chart specified in MIL-STD-150A is somewhat dated but it allows one to readily obtain a basic evaluation of the spatial resolution performance of a camera system.

The other widely used tool to evaluate spatial resolution performance in humans is the optometrist's logMAR chart[1] The logMAR chart was devised by Bailey and Lovie [1]. It consists of geometrically scaled lines of lettering in the Sloan font. The rows of letters on the chart are scaled logarithmically, each row being scaled $10^{0.1}$ relative to the last. This means that the letter height approximately doubles every third row. The chart is constructed on the basis that a person with normal sight can resolve 1 minute of arc, this is the Minimum Angular Resolution (MAR). The chart is typically designed to be viewed at a distance of 6m (20 feet) so letters at the bottom of the chart with a logMAR value of 0 are sized to match this at approximately 9mm high, see Figure 5. This row of the logMAR chart is marked by the single horizontal line across the bottom of the chart seen in Figure 6. Subsequent rows of letters above this are numbered by logMAR values increasing by 0.1 at each line to a value at the top of 1. The logMAR 0.5 row is marked by the double horizontal line across the chart. The letters in this row are $10^{0.5} \approx 3.16$ times larger than the logMAR 0 row, about 27mm high.

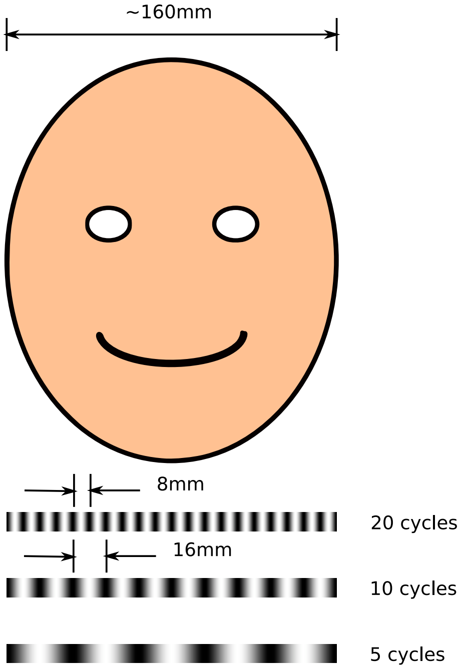

Figure 3. Spatial frequencies important for human face recognition.

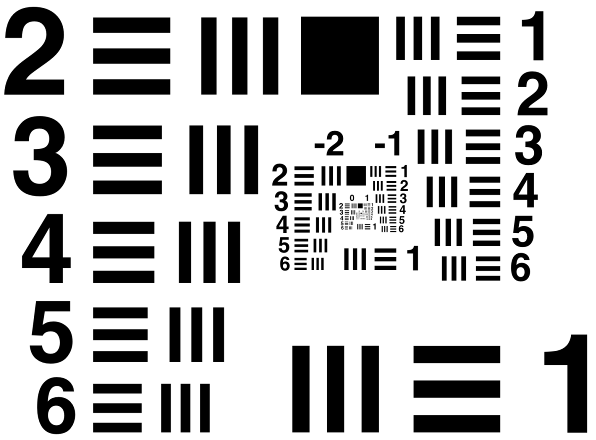

Figure 4. 1951 USAF Chart, composed of groupings of 6 geometrically scaled pairs of bars. Each successive grouping is half the size of the previous. The bottom right bars have a spatial wavelength of 16mm, the top right have a wavelength of 8mm. These encompass the spatial frequency range important for face recognition

Visual acuity is also often described in terms of the Snellen fraction; the ratio of the distance you can read a specific line of the eye chart to the distance that someone with normal vision would be able read that line. Thus the term 20/20 vision (or in metric, 6/6 vision) arises from the fact that at 20 feet you can read the line on the eye chart that is legible to someone with normal vision at 20 feet. If you can only read the line at logMAR 0.3 (letters twice

[1] Here logMAR stands for 'logarithm (to base 10) of Minimum Angular Resolution', it has nothing to do with the Laplacian of Gaussian and David Marr!

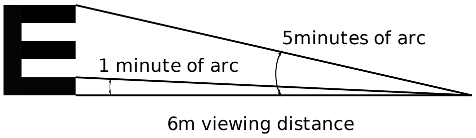

Figure 5. A person with normal sight can resolve 1 minute of arc (logMAR 0). Letters on the 6/6 row of an eye chart will be sized to match this at just under 9mm high.

the size) your Snellen fraction would be 6/12 (at 6m you can only read something that should be visible from 12m). The letters at the top of the chart, which are approximately 90mm high, correspond to a Snellen fraction of 6/60. This represents 1/10 of normal visual acuity. If you are unable to read this top row you would be considered to have a 'severe visual impairment' and be legally blind [21]. See Figures 13 and 14.

While the logMAR chart is certainly not the most precise instrument one would choose to evaluate the performance of a camera it is useful in that it provides a direct comparison with personal experience. It allows one to readily communicate to non-technical people an important aspect of the performance of a surveillance system. One can relate a 6 metre logMAR chart to the USAF chart by noting that the key spatial frequencies with wavelengths of 8mm and 16mm, corresponding to the top-right and bottom-right bars of the USAF chart, roughly match the logMAR values of 0.4 and 0.7 respectively. (The 6m logMAR values of 0.4 and 0.7 actually correspond to spatial wavelengths of 8.8mm and 17.4mm respectively.)

III. Experiments

In evaluating a surveillance system's visual acuity we are not necessarily trying to determine whether it has 'normal human vision'. All we want to know is whether it can resolve the spatial frequencies that are important for face recognition at the operating distance the camera is working at. To test this we can image the test charts at the camera's operating distance, which might be much less than 6m, and check whether the appropriate USAF bars or logMAR lines can be resolved. Nevertheless in the interest of doing a simple comparison against human vision some basic experiments were conducted with a Pulnix TM6CN 1/2" CCD camera positioned 6m from the USAF and logMAR charts. Four life size face images were also placed alongside the charts. Using C-mount lenses ranging in focal length from 4mm to 16mm a variety of images were digitized using a Data Translation 3155 frame grabber. The results are shown in Figures 7 to 10.

Naturally due to the size constraints of images in this paper and variations in image reproduction it is hard to provide a proper evaluation of these images. However, it is hoped they provide a useful indication of the general trends.

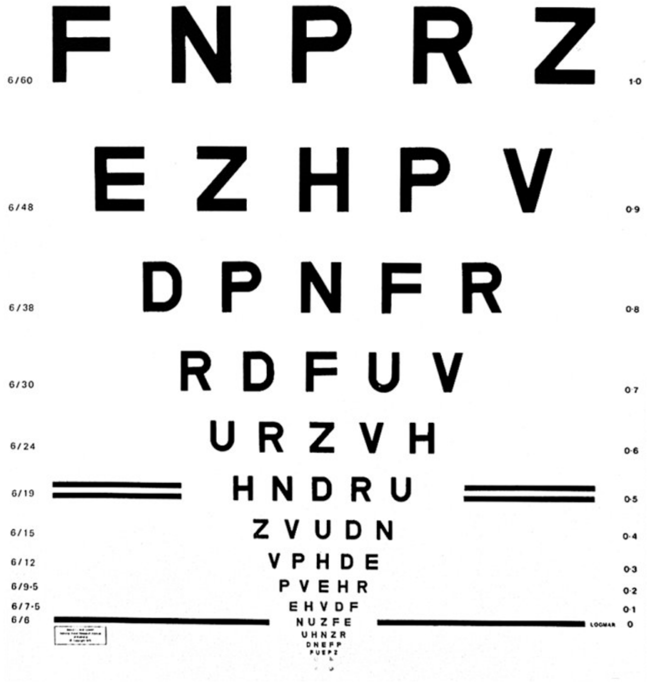

Figure 6. The logMAR chart, letter heights approximately double in size every third row.

Figure 7 shows that the image from the 4mm lens is close to satisfying the definition of being legally blind. The top row of the logMAR chart is only just legible. Looking at the USAF chart and also noting that the key spatial frequencies for face recognition lie between logMAR 0.4 ('Z V U D N') and 0.7 ('R D F U V'), we can see that only when one uses a 12.5mm lens (Figure 9) do we have sufficient image quality to recognize a face with some confidence.

The requirement of being able to resolve spatial frequencies of 20 cycles per face width would suggest that a face should be at least 40 pixels wide in the image for one to be able to recognize it with some reliability. Figure 9 would appear to support this view as the heads in this image are close to this value at about 38 pixels wide. However this presupposes that there are no other deficiencies in the image quality. Note that the head widths in Figure 1 and in the two images shown in Figure 14 are approximately 35 pixels, 30 pixels and 20 pixels respectively.

IV. Compression

These images presented in Figures 7 to 10 represent a 'gold standard' quality in that they are acquired with a relatively good quality camera and lenses, and directly digitized from the camera with a good quality frame grabber. In practice many systems would employ cheaper cameras and lenses and the image would be recorded to video tape, or compressed and stored digitally.

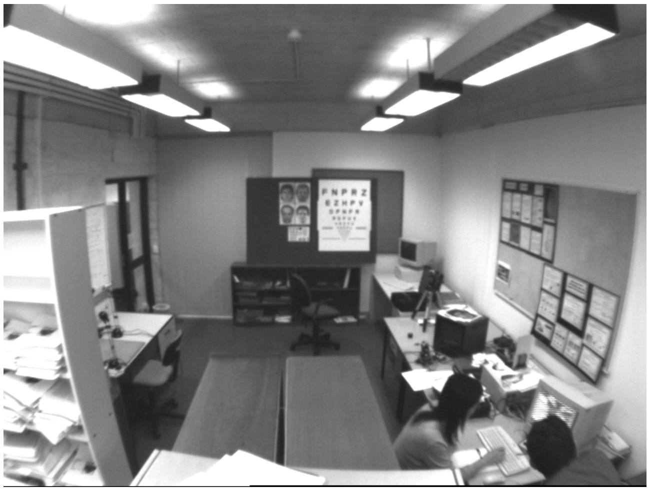

Figure 7. Image using a 4mm lens.

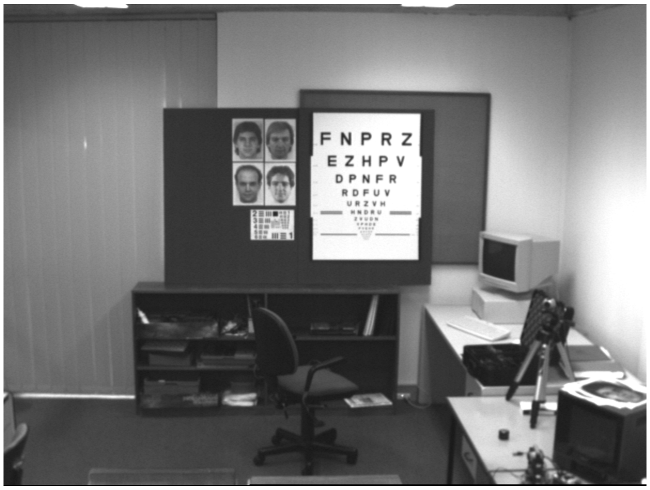

Figure 8. Image using a 8.5mm lens.

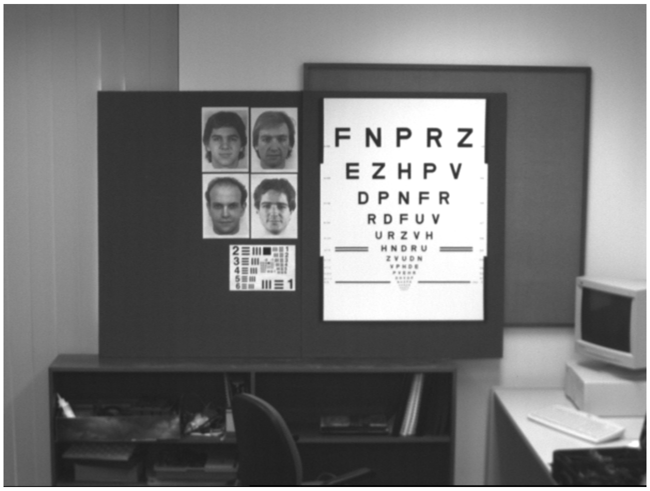

Figure 9. Image using a 12.5mm lens.

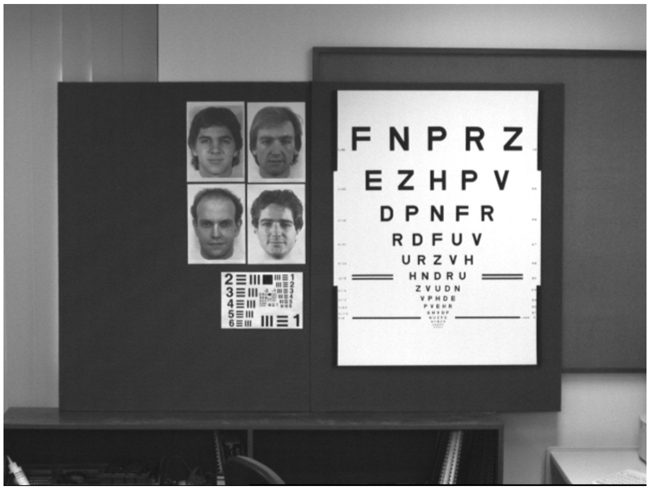

Figure 10. Image using a 16mm lens.

Figure 11 shows a closeup of the original 12.5mm lens image (a), a version of the image recorded to video and replayed and digitized (b), and two compressed versions of the original image (c) and (d). Inspection of the vertical and horizontal bars of the USAF chart in the image that was recorded to video would indicate that while vertical resolution is mostly unchanged, horizontal resolution is approximately halved. This would greatly reduce the confidence with which one could identify a face.

However, it is compression that is potentially most problematic for surveillance images. Often digital surveillance systems will use quite aggressive compression ratios because of the amount of image data that has to be stored. Figures 11(c) and (d) show that while simple test grating patterns, such as the USAF target used here, survive data compression quite well, faces can be degraded considerably.

At first sight this is at odds with what has been reported elsewhere. For example in the FRVT 2006 test it was reported that the low resolution images were compressed 20:1, yet excellent recognition performance was achieved. Griffin and Hsu[12] do report a clear degradation of automated face recognition with compression ratios beyond 20:1. However, below this ratio the effect on performance is negligible. Delac et al. [7] review the automated face recognition literature finding numerous studies indicating that compression ratios up to around 20:1 have little effect. The FRVT 2000 evaluation report [9] even suggested that ratios up to 40:1 had little impact.

I argue that these results indicating that compression has little effect on recognition performance are not relevant to surveillance images because they relate to images having much greater resolution. The images in the low resolution FRVT 2006 test were such that the distance between the eyes was about 75 pixels. Using an average interpupillary distance of 63mm [8] and an average head width of 160mm

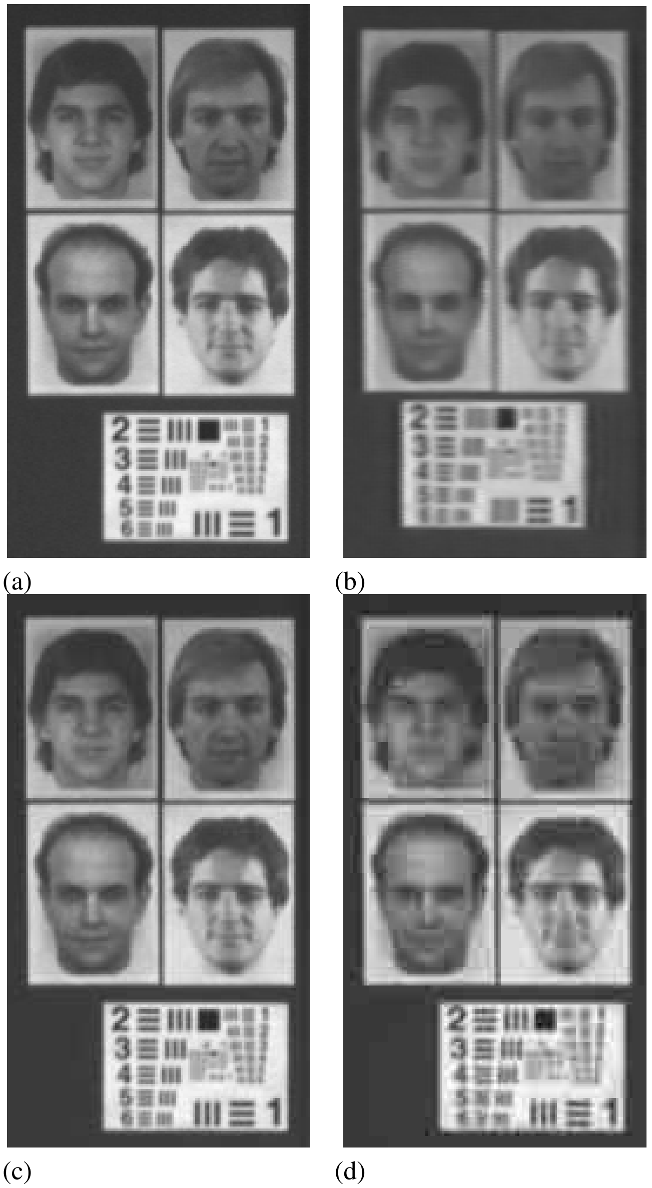

Figure 11. Close up of face images and USAF target recorded with the 12.5mm lens. (a) Original image. (b) Image recorded to video tape, replayed and digitized. (c) Original image with JPEG compression 18:1. (d) Original image compressed 31:1.

we can estimate the width of the head to be about 190 pixels. In the results reported by Griffin and Hsu [12] the distances between the eyes were 120 pixels, which translate to a head width of about 305 pixels. In comparison the faces shown in Figure 11 are approximately 38 pixels wide.

In this case these images are only just short of the Nyquist limit of providing spatial information at 20 cycles over the face width. However, the main problem is the interaction of the 8 × 8 MPEG or JPEG discrete cosine transform (DCT) blocks with the image. There are approximately 5 of these DCT blocks spanning each face. The compression quantization of the frequency components in each of these blocks means that the spatial frequency information from 5 cycles per face width upwards have been corrupted. This is precisely the range of spatial frequencies important for face recognition! The discontinuities introduced at the block boundaries are especially troublesome. In the limit we have all the perceptual problems associated with block masking [13]. This can be seen with the excessive compression shown in Figure 12. Thus, at the typical resolution of surveillance images, compression is likely to have a very significant effect on recognition performance.

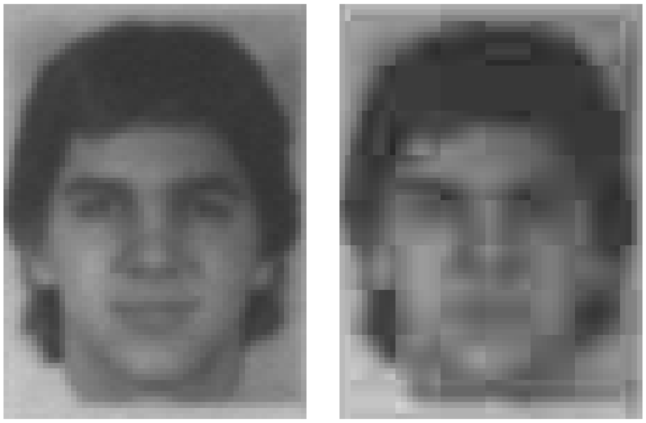

Figure 12. Further enlargement of uncompressed and 31:1 compressed face from figure 11 showing JPEG DCT blocks.

In comparison the low resolution FRVT 2006 images, where the face width was about 190 pixels, would have been hardly affected by the JPEG DCT blocks. It takes about 23 8 × 8 blocks to span these faces. Quantization of frequency components in these blocks will have little impact on human perception as these are beyond the important range for recognition. This may well explain the observations that image compression has little effect on automated face recognition. In saying this, one must acknowledge that the spatial frequencies that are important for human recognition of faces are not necessarily the same as those that are important for automated recognition systems. However, it is likely there is a strong relation between the two.

V. Luminance and Colour Information

While the discussion so far has concentrated on the spatial frequencies important for face recognition and the disruption of them by the compression process, luminance and colour cues are at least as important as shape cues. O'Toole et al. [18] and Russell et al. [19] demonstrate that people perform about equally well using either just shape information, or just pigmentation information to recognise faces. Referring to Figure 15 one can see that the faces in the top and middle rows, which differ in only shape or pigmentation respectively, do not look the same as each other indicating that both shape and pigmentation are important.

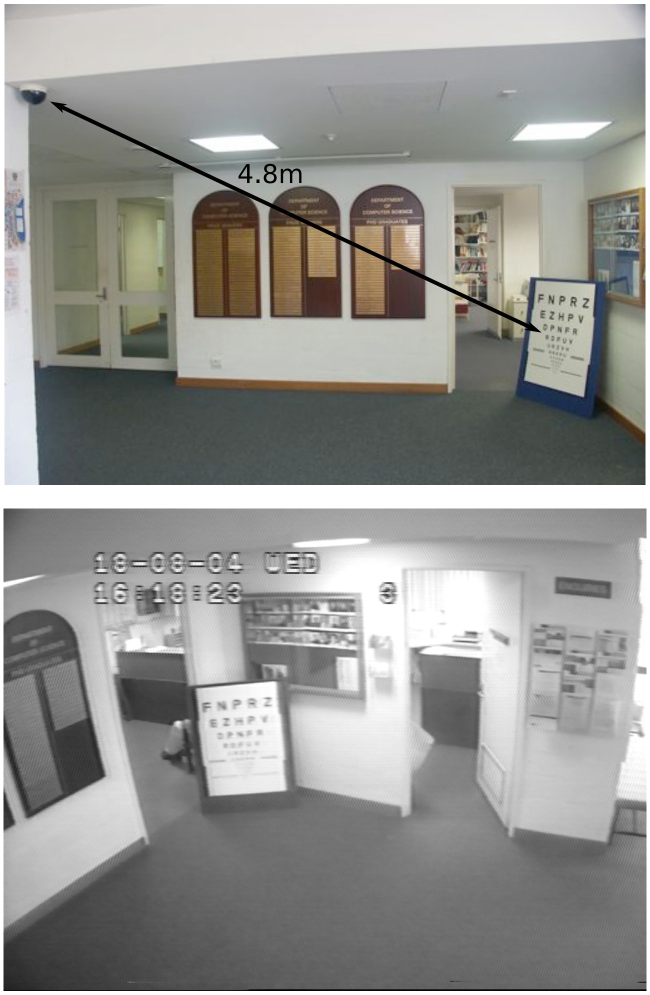

Figure 13. An indoor camera installation and its image. Note the distance from the camera to the logMAR chart is only 4.8m.

The importance of pigmentation is of concern given that under MPEG there are the additional quantization issues introduced by the 16×16 motion prediction macroblocks. While luminance information is encoded within the four 8×8 DCT blocks within a macroblock the chrominance information is typically subsampled by a factor of two and is thus encoded at the macroblock size. The DCT values representing chrominance information are also more heavily quantized than those representing luminance data. This is illustrated in Figure 16.

VI. Conclusions

Surveillance cameras, as they are currently used, are almost useless for the identification of people. The poor resolution of surveillance images means that it is likely that many of the spatial frequencies important for face recognition will not be present. In addition, the low resolution means that compression DCT blocks will interact strongly with the spatial frequencies important for face recognition.

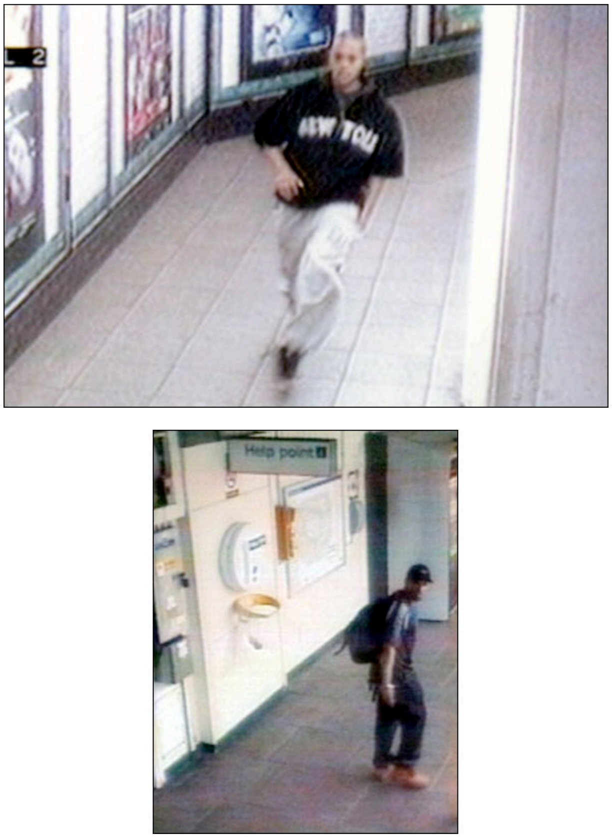

Figure 14. Images released in relation to the attempted bombings in London on July 21 2005 [2]. In the top image the letters on the jacket are about 80-90mm high, assuming the head is 160mm wide. They are not legible. In the second image the 'Help Point' letters on the overhead sign are larger still, yet only just legible. It might be argued that these surveillance cameras are legally blind!

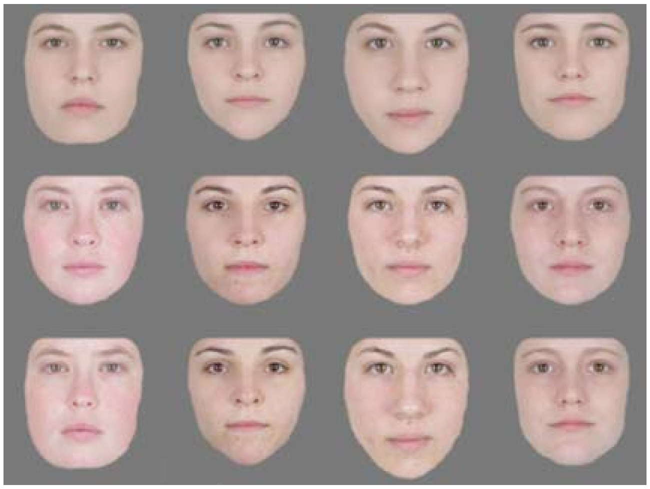

Figure 15. Illustration from Russell et al. [19]. Faces in the bottom row are images of laser scanned faces differing in both shape and pigmentation. Faces in the middle row differ only in pigmentation, not shape. Faces in the top row differ in shape but not pigmentation.

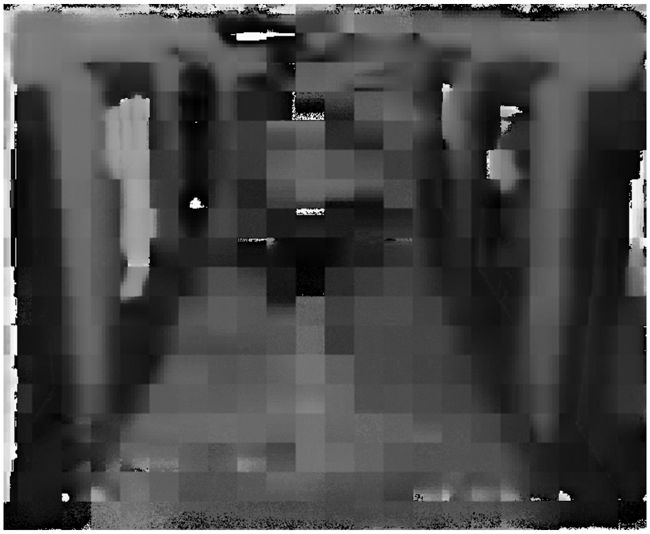

Figure 16. Hue values extracted from the image shown in Figure 1 displayed as greyscale. Where is the person?

Finally, the quantization of colour information introduced by compression adds a further confounding of our ability to recognise faces in surveillance images. These conclusion are illustrated by images from surveillance cameras that arguably meet the definition of being legally blind.

References

[1] Bailey IL, Lovie JE. “New design principles for visual acuity letter charts”. *Am J Optom Physiol Opt* 53: 740-745, 1976.

[2] “Police issue bomb suspect images” BBC news website. Friday, 22 July, 2005. http://news.bbc.co.uk/2/hi/uk_news/4706421.stm

[3] Vicki Bruce, Zoë Henderson, Karen Greenwood, Peter J. B. Hancock, A. Mike Burton and Paul Miller. “Verification of face identities from images captured on video”. *Journal of Experimental Psychology: Applied*. Vol 5(4), Dec 1999.

[4] Vicki Bruce, Zoë Henderson, Craig Newman, A. Mike Burton. “Matching Identities of Familiar and Unfamiliar Faces Caught on CCTV Images”. *Journal of Experimental Psychology: Applied*. Vol 7(3), pp207–218. 2001.

[5] Burton, A.M, Wilson, S., Cowan, M and Bruce, V. (1999). “Face recognition in poor quality video: evidence from security surveillance”. *Psychological Science*, 10, 243-248

[6] Costen, N.P., Parker, D.M. and Craw, I. “Effects of high-pass and low-pass spatial filtering on face identification”. *Perception and Psychophysics*, 58, 602–612. 1996

[7] Kresimir Delac, Sonja Grgic and Mislav Grgic. “Image Compression in Face Recognition - a Literature Survey”. in *Recent Advances in Face Recognition*, edited by: Kresimir Delac, Mislav Grgic and Marian Stewart Bartlett, pp. 236, December 2008, I-Tech, Vienna, Austria.

[8] Neil A. Dodgson, “Variation and extrema of human interpupillary distance”. *Proc. SPIE Vol. 5291 Stereoscopic Displays and Virtual Reality XI*. San Jose. pp36–46. 2004

[9] Blackburn D.M., Bone J.M., Phillips P.J., “FRVT 2000 Evaluation Report, 2001”, available at: http://www.frvt.org/FRVT2000/documents.htm

[10] P. Jonathon Phillips, Patrick Grother, Ross J. Micheals, Duane M. Blackburn, Elham Tabassi, and Mike Bone. “Face Recognition Vendor Test 2002”. NISTIR 6965, March 2003 , available at: http://www.frvt.org/FRVT2002/documents.htm

[11] P. Jonathon Phillips, W. Todd Scruggs, Alice J. OToole, Patrick J. Flynn, Kevin W. Bowyer, Cathy L. Schott, and Matthew Sharpe. “FRVT 2006 Report: FRVT 2006 and ICE 2006 Large-Scale Results”. NISTIR 7408, March 2007, available at: http://www.frvt.org/FRVT2006/Results.aspx

[12] Paul Griffin and Vincent Hsu, “JPEG and JPEG2000 Compression for Face Recognition” available at: http://fingerprint.nist.gov/standard/archived_workshops/ workshop1/presentations/Griffin-Face-Comp.pdf April 2005.

[13] Harmon LD and Julesz B, “Masking in visual recognition: Effects of two-dimensional filtered noise”. *Science* 180:1194–1197. 1973.

[14] Hayes, T., Morrone, M.C, and Burr, D.C. “Recognition of positive and negative bandpass-filtered images”. *Perception*, 15, 595–602. 1986.

[15] Richard Kemp, Nicola Towell and Graham Pike. “When Seeing Should not be Believing: Photographs, Credit Cards and Fraud”. *Applied Cognitive Psychology*. Vol. 11 pp211–222. 1997.

[16] Peter Kovesi, ”Phase Preserving Denoising of Images”. *The Australian Pattern Recognition Society Conference: DICTA’99*. Perth WA. pp 212-217. 1999

[17] Risto Näsänen. “Spatial frequency bandwidth used in the recognition of facial images”. *Vision Research*. 39 (1999) 3824–3833.

[18] A. J. O’Toole, T. Vetter, and V. Blanz, “Three-dimensional shape and two-dimensional surface reflectance contributions to face recognition: An application of three-dimensional morphing”, *Vision Research*, vol. 39, pp. 31453155, 1999.

[19] R. Russell, P. Sinha, I. Biederman, and M. Nederhouser, B. “The utility of surface reflectance for the recognition of upright and inverted faces”. *Vision Research* Volume 47, Issue 2, January 2007, Pages 157-165.

[20] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell, “Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About”. *Proceedings of the IEEE*. Vol. 94, No. 11, November 2006.

[21] “Visual Standards: Aspects and Ranges of Vision Loss with Emphasis on Population Surveys”. Report prepared for the International Council of Ophthalmology at the *29th International Congress of Ophthalmology*, Sydney, Australia, April 2002.