Draft framework for the intended Australian CCTV Standard 4806.5 referring to Digital and Networking in CCTV

(draft prepared by Vlado Damjanovski, June 2008)

Note: *The following text is only a loose proposal for the intended AS 4806.5 standard and is a result of an ongoing on-line collaboration between interested members and non-members of the Australian CCTV industry. The medium for this collaboration is the standards forum discussion "(DRAFT) - Digital and Networking in CCTV" proposed standard forum at* http://www.el51televisionstandards.com.au/cctv.html
which is moderated by Les Simmonds and Vlado Damjanovski. All of the text published on this forum site remains copyrighted property of the moderators until it is finalised and brought up as a final document by Standards Australia.
.

FOREWORD

This Standard will prove useful to those responsible for establishing operational requirements, writing specifications, selecting equipment, installing, commissioning, using and maintaining a CCTV system with digital video signals.

Closed circuit television, in its simplest form, is a means of providing images from a television camera for viewing on a monitor via a transmission system, which can be in analogue or digital format.

This standard is in addition to the AS4806.1, AS4806.2, AS4806.3 and AS4806.4 and should be read in conjunction with them.

There is no theoretical limit to the number of cameras and monitors that may be used in a CCTV surveillance installation but, in practice, this will be limited by the efficient combination of control and display equipment and the operator's ability to manage such system. Furthermore, with the introduction of digitally encoded signals, the closed circuit environment becomes a much wider circuit using local area networks (LAN), wide area networks (WAN) and the Internet as a medium.

Scope

This Standard provides recommendations for the selection, planning and installation of closed circuit television systems comprising camera(s), encoder(s), streamer(s), network switches, decoder(s), monitor(s), video recorder(s), switching, control and ancillary equipment for use in security and surveillance applications.

NOTE: This Standard should be read in cinjunction with the AS4806.2.

1 The "Digital CCTV" standards

At the time of preparing this standard the CCTV industry still uses analogue CCTV cameras, some use digital CCTV cameras, but also encoders that convert analogue signals into digital and a mixture of all of the previous.

For this reason the Digital CCTV standards in this document should be treated only as an addition to the existing Australian CCTV standards. The global recommendations and design criteria of putting a CCTV system together, as described in the AS4806.1, .2 and .3 are still valid and are complementary to this (AS 4806.5) document.

The subject of AS4806.5 are the so-called “IP CCTV cameras,” variety of compression, networking of CCTV and other specifics that are coming out from the digital signal processing, transmission and storage.

Digital CCTV can be described as a range of products, procedures, codecs, transmission and storage devices where CCTV information is represented with digital signals.

2 *Definition of a digital CCTV signal*

A digital CCTV signal is considered any image or video signal of a CCTV camera that is represented in digital format (with “0” and “1”), using bits and bytes. The most common digital representation of an analogue CCTV camera image is the luminance and colour of each camera picture element (pixel) with 8-bit digital numbers (256 variations per colour), representing the equivalent analogue value of each pixel luminance and colour.
The number of pixels in such a digitised image is typically equivalent to the number of active pixels in the imaging device producing such an image .
For a standard definition CCTV camera, the pixel count is defined by the imaging chip manufacturer, and for PAL, this is typically 752x582, equating around 440,000 active pixels. We will consider this as a Standard Definition CCTV image. In the analogue world such an image, with good quality lens is expected to produce an image with a horisontal resolution of around 480TVL, which is equivalent to around 6MHz of analogue signal bandwidth.

As defined in AS4806.2, in Australian CCTV we use the ITU-601 compliant digitisation process, which states that 13.5MHz sampling is used to digitise the analogue video signal of each line of the video signal produced by the imaging chip made up of the above mentioned 752 pixels per line. This produces 720 samples (not 752) of each of the 576 active lines, hence we will consider that digitised standard definition CCTV frame is composed of 720x576 pixels, making a total number of 414,000 active pixels. This many refer to as 4CIF resolution, or in the American industry as D1.

After the digitisation of the video signal, we use compression.

The compression can be *image compression* when a single individual TV frame or TV field is compressed, or a *video compression* when sequence of TV fields or frames are compressed.

This standard will discuss the variety of compression and what the CCTV industry expects from each one of them, but the most important parameter we would like to introduce with this standard here are the qualification of the image quality from any device, whether that be a standard definition live streaming or megapixel camera with a specific lens, even a “non-standard” imaging chip and proprietary or unknown compression.

Introduction of Identification CCTV Unit (ICU)

In order to lay a solid foundation that will be universal and applicable to any type of digital CCTV video signal we must introduce some universal image quality quantification methodology. This should be a methodology that could be relatively easy aapplied on any system, analogue or digital, standard definition, high definition or megapixel.

The idea of this methodology is something that an be easily used by an average installer, consultant and, of course, the end user.

There are a number of parameters that needs to be considered in a system, and all of the following influence the end result, so they all must be included:

- Angle of view
- Lens quality
- Imaging chip pixel count
- Camera digitisation and processing quality
- Images per second captured, transmitted and/or recorded
- Latency (including digitisation, encoding and network)

The following headings and paragraphs are pillars of the new Digital CCTV Standards.

1 The key parameters of a digital CCTV signal quality

1 Pixel count

1. This is defined as physical image size in active picture elements (pixels).
   A standard definition (SD) analogue PAL signal, when converted to a digital signal is usually compliant with the ITU-601 sampling recommendation, producing an image of 720x576 pixels when two TV fields are put together. Some times, in CCTV, we refer to these as “4CIF” resolution, or “Full PAL TV frame”, or in some American literature – as “D1” resolution.
2. High Definition (HD) TV is defined as 1920 x 1080 pixels.
3. Many new cameras and imaging chips are now available on the market, often referred to as “Megapixel” cameras. These are cameras with imaging chip producing over 1 megapixels, and most common 2 megapixels (such as HD 1920x1080=2,073,600 pixels), but 3, 5, 8 or even 11 megapixels imaging chips cameras are also available.
4. The important thing to note is also that each of these imaging chips may have different imaging chip size, ranging from 1/4” up to 2/3” or even larger. This then, influences the individual pixel size, which in turn affects the noise and minimum illumination performance. With different imaging chip sizes the angle of coverage and lens characteristics may be different. This influences the overall quality.

2 Images per second

1. All of the above “images” can be produced (and subsequently captured) at a rate defined by the imaging chip and the camera electronics. In Australia we consider 25 i/s as “live rate”. This is equivalent to images taken at 40ms intervals. Using the most common analogue interlaced concept, we produce TV frames of each two fields, which are captured at 20ms intervals, but when digitally superimposed, they are 40ms “apart”.

2. Many HD cameras can capture and produce 25 i/s or less, depending on the design.
3. Megapixel cameras, especially the ones with more megapixels than HD, with today technology usually capture less images/second than at “live rate”. This is clearly defined by the imaging chip and the camera electronics. Since in CCTV we have quite a variety of such “non-live-HD” megapixel cameras it is important to know not only how many megapixels a camera has, but also how many it can capture, process (encode) and produce at the output.
4. We must introduce here a statistcial and experimentally proven “human reaction” response time of 200ms. Namely it has been proven that human reflexes are not faster than 200ms (an average figure). This indicates that when CCTV is used in “normal” surveillance of human activity (e.g. if somebody is attacked in a shopping centre, shoplifting, falling, etc.) 200ms, or 5 i.s may be considered sufficient. Clearly, in systems such are casinos, or cash handling areas, where cameras usually have narrower angle of view, motion of hands is faster in appearance relative to the image projected on the camera chip, and more than 5 i/s might be required. This is a general observation to indicate that we must take into account i/s base don human activity.

3 Lens quality and angles of viewing

1. No matter how many megapixels a camera has, it is quite clear that the lens quality must be of adequate quality if we would to achieve the expected image quality. This usually means the lens should have at least the same resolving power as the pixel density, but preferably more. One lens that might have sufficient resolving power for one imaging chip might not be suitable for another where pixels are smaller.
2. The angle of vision is another important parameter. Even with inferior lens, if we have a tight shot of an intruder's head, we may still produce a meaningful analysis. The opposite is true too, even with the best lens, if the angle of vision is too wide, we may have intruders illegible as they may appear too far from the camera.
3. Correctly focused lens on an object can also be introduced here as an important factor, especially with megapixels camera where “live” view is not always available for interactive focus adjustment.

4 Video/Image compression

1. All digital (IP) cameras have analogue to digital conversion circuitry inside, followed by a compression (either Image or Video compression) stage. All compressions in CCTV are lossy, which means there is doubt there is a further picture quality loss, the question only is by how much.
2. Since most compression levels can be configured freely, it means that image quality can be further reduced (after being digitised) by the compression level (not only by the type), so that we have to introduce a parameter that will quantify the image quality loss due to compression.

5 Latency

1. Image capturing, digitising and compressing it, as well as transferring it via the network, does take some time. This can range anything from a few milliseconds up to few seconds. Some systems may extend larger areas, territories or even continents, so clearly this might be a problem for interactive and “live” control of such systems. We have to introduce some measurement that will encompass these parameters. Again, as under the “Images per second” heading, we should consider the “human reflex of 200ms” as a reference point.

2 The starting point for unity quality (1 ICU)

1 Definition of 1 ICU reference (ICUr)

The definition of Face Identification in AS4806.2 is a very natural and logical starting point for the quantifying unity quality.

If we take 100% of average human person to be in the field of view of a camera, filling the monitor height, we can derive from it the following:

In Australian PAL CCTV we have 576 active lines, which when converted to digital (before compression) give us 576 pixels in vertical direction. An average person's head takes around 15% of it's own height, which converted to pixels makes it around 86~88 pixels.
Let's use 88 pixels as it is easy to remember and allows for small error in angles of view, variation in human head sizes, etc.

We could consider such an image of a person (using a good lens, properly focused camera, and before the image is being compressed) as a reference unity quality for face identification, based on AS4806.2.

We call this “Unit of face identification based on Australian CCTV standards”, or shortened “Identification CCTV Unit” which we will refer to further in this document as ***ICU*** (“I see you”) for easy remembering and symbolic reference.

If we have for example a system that is designed, installed and adjusted to cover certain areas where people would be captured in their full height filling the camera imaging chip height (as seen on a 400TVL monitor, as defined in the AS4806.2) we would call such a system that has value 1 of ICUs.

2 Variation from 1 ICU

Let's say now we have a similar system, but the end result is only a CIF size image (288 vertical pixels), but it also shows a person in its full height. Because you would have only half of the required pixels for unity of ICU, we could quantify this as having 0.5 ICU.

If a system with CIF size image is used, in order to have 1 ICU, the angle of coverage has to be such that the persons head occupies around 88 pixels height, which is equivalent to having a view of a standing person filling the whole screen height (288 pixels) up to his waist (200% person's height – using the same logic as explained in AS 4806.2).

Another example, if we have a 2.3 megapixel camera (with for example 2000 x1152 pixels) that has a good lens (well focused) and it also covers a full persons height in one shot. This camera would have 2 units of ICU (2 x 576 = 1152), if the lens is positioned to have such an angle of vision.
Since this is a megapixel camera, it would be non-productive to have such a “narrow” view for the purposes of face identification, so we would naturally choose a wider angle of view so that we see a person in half of the image height, qualifying such a camera to have unity (1) ICU.

Using the same logic, if we would to use a CIF recording system, to produce unity ICU, we would need to use angle of view that will cover 50% of a persons height.

Using the same logic further, we could quantify a camera in a system that is encoded and recorded with high quality 4CIF resolution, but the angle of coverage of the lens is wider than what is required for face identification. This is certainly not a problem if such a camera is intended to cover a wide shot of the hotel foyer for example, but using the above logic we could still quantify this view, which includes the pixel count as well. So, if we assume that the view is wide and shows a persons height to be only half od the monitor screen, we could still refer to this as camera with 0.5 ICU units.

There is an easy way to convert ICU units to other values specified by the standards using simple mathematics. For example, the AS 4806.2 recommendation for visual recognition of vehicle number plates where it is suggested that number plate characters should be at least 5% of the image height. This would mean that at the image plane where the number plate is located, the vertical height of such a TV frame would be around 1.4 m. this is obtained from the assumption that an average number plate character is 70mm (in Australia there are characters from 60mm up to 80mm), which would then be 5% of the image height equal to 1.4m. Such a view (when the vertical height is 1.4m) is narrower than when the vertical height is equal to seeing an average person (1.7~1.8m) which is the view recommended for 1 ICU. So we get that 1.7/1.4 = 1.2.
The conculsion would be that if a camera has a view to satisfy the number plate visual recognition (characters to be 5% of the image height) such a view would produce a 1.2 ICU. In other words, it would be more than sufficient for an operator to make facial identification, as per AS 4806.2.
Using similar logic, playing cards in a casino can be identified visually when the cards size is at least 10% of the image height (60 pixels minimum). Considering a standard playing card is 9 cm tall, the filed of view where a card height is 10% is equal to 90cm image heigth. At this image height we would have half a person in the field of view, which means 2 ICU are required for playing cards visual recognition (assuming no deterioration from the lens and compression occur).

3 Introducing normalisation parameters for images/second (ICUi)

The above points 1 and 2 are only introduction of unity ICU definition but without consideration of images/second. So, clearly, the best we would hope to get in digitised CCTV would be 25 i/s. It is therefore suggested that unity ICU to indicate that we have 25 i/s of a full 100% persons height in a SD CCTV (before compression).

So, if we use, for example, a camera (or system) that produces only 5 i/s, we should multiply the unity by a percentage so that it indicates that the ICU quantification for such a system is less than 1.

This document introduces here a non-linear “normalisation.”

As mentioned previously, in most “normal” CCTV systems 5 i/s is still quite good, although not as good as 25 i/s. The “almost unacceptable” or “just sufficient” rate is considered to be 1 i/s. So, if we refer to 25 i/s as the maximum I can get and call it “unity”, the 5 i/s is not calculated as 1/5 (20% or 0.2) of the unity, but rather a 50% value is suggested (0.5).

For 1 i/s I would rate is as 10% of unity (0.1).

Here is a table suggestion based on practical experience:

25 i/s = 100% (unity = 1)
20 i/s = 90% (0.9)
15 i/s = 80% (0.8)
10 i/s = 70% (0.7)
5 i/s = 50% (0.5)
3 i/s = 30% (0.3)
2 i/s = 20% (0.2)
1 i/s = 10% (0.1)

So, for the above mentioned examples, if we have a 4CIF digitised signal, at 25 i/s, this would have unity ICUs.

If the same is referring to 5 i/s I would quantify it as having 0.5 ICUs.

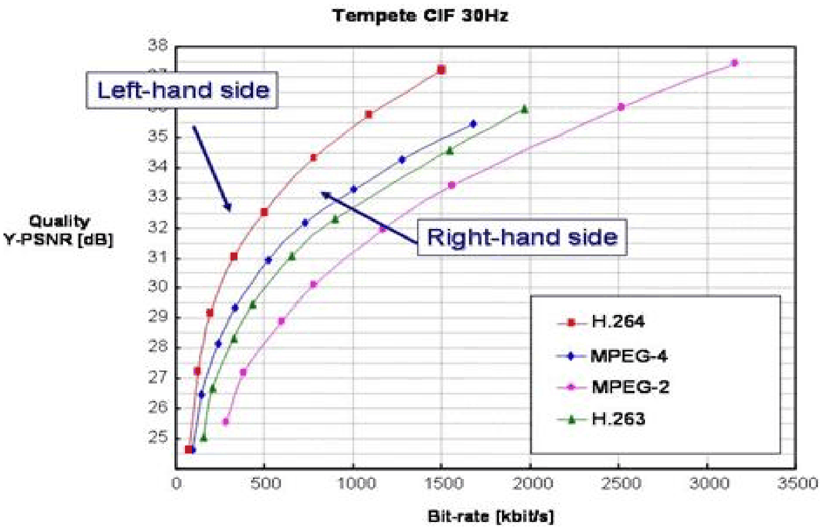
The 2.3 megapixel camera mentioned previously, if being set-up to produce 1 ICUs before the introduction the i/s parameter (two persons heights in the filed of view) then having 5 i/s will obtain 0.5 ICU (this time with i/s weighing factor). If the angle of view is covering a full persons height, that is 2 ICUs, than using 5 i/s this will give us 1 ICU overall.

4 Introducing the compression normalisation parameter (ICUc)

All of the above are recommendation which are based on the image produced by the camera before the image gets digitised and compressed.
It is quite clear though, that the image or video compressions used in CCTV are lossy compressions, and as such do introduce loss of details.

How do we measure loss of details in compressed image or video?

This is very difficult to measure, but fortunately there are objective methods for exactly this – called Peak-to-signal-noise ratio” (PSNR).
There are many measurements and tables available for various compression depending on the streaming rate and the type of compression.
Ultimately, the CCTV industry should suggest and/or develop independent bodies to quantify such PSNR values for all compression available on the market.



For illustration purposes only, we could use the following graph:
Let's take 1 Mb/s in the above and we will see MPEG2 has a PSNR of

around 31.3dB, while for the same streaming rate, H264 produces almost 35.3dB (4 dB difference). We are aware from general electronics and definition of decibels (dB) that each 3 dB is approximate 50% difference in the ratio. The example above might be a bit awkward, but 4dB is approximately 60% difference (20log160=44dB, 20log100=40dB). So we could say that at 1 Mb/s the H264 compressed signal will visually produce a signal quality that is approximately 60% better than MPEG2 at the same rate.
In our example above if we use 1Mb/s of MPEG-2 compression, so we could say that this video quality is around 62% (160/100) of the same encoded with H264.

But, which compression should we use as a reference unity ICU compression factor?

The Australian CCTV Standard committee suggests to start from a known parameter for SD TV, and that is the DVD quality which is known to the general public. The DVD uses MPEG2 video compression with compression of around 4 Mb/s. The above graph doesn't show MPEG-2 up to 4Mb/s, but if we assume we know it, we can safely guess (of course this will be confirmed for the final document) that MPEG-2 PSNR is be around 38dB, the same PSNR for H263 would be around 39dB, for MPEG4 around 40dB and for H264 around 43dB.

If the above is a correct, than we could say that we have 1 ICU video quality if we have a SD camera covering an area with a person height filling the image height fully (100%), being encoded and recorded at 25 i/s and using MPEG-2 encoding of 4 Mb/s (PSNR=39dB). If the same video sequence is encoded with H-264 using a 2Mb/s it would produce the same PSNR (video quality, according to the graph), e.g. this would be “unity ICUs”.

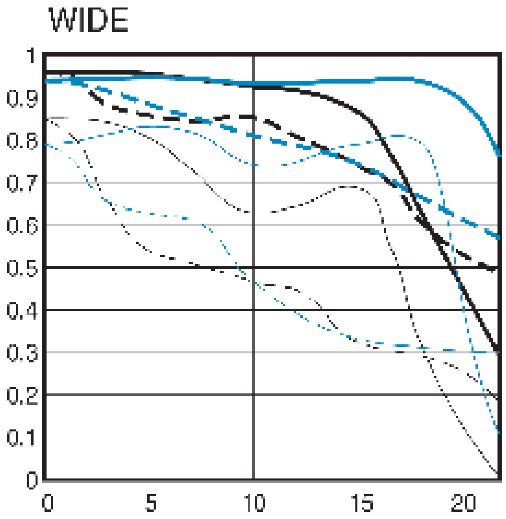
If we have another example where the PSNR is found to be 35dB (that is 3dB less than 38dB, representing about 67% of the 38dB), then for a full PAL video, covering a full persons height (25i/s), and assuming we know such a compression has PSNR of 35dB would produce an ICU of 0.67.

5 Introducing the lens normalisation parameter (ICUo)

This is perhaps the easiest to understand, but hardest to measure and quantify. The key behind the lens normalisation parameter which will indicate the lens quality, is the Modulation Transfer Function (MTF).

Standards Australia may suggest sponsoring or endorsing organisations (like CSIRO for example, or private specialised companies) that will perform independent and verifiable testing of CCTV lenses, producing MTF curves for each lens.
The MTF characteristics for lenses should show the resolution power in line pairs per millimeter (lp/mm). Line pairs are used because for an imaging chip to produce distinguishable lines, both black and white lines are needed, hence line pairs. Since there are a variety of imaging chips sizes (e.g. 1/4”, 1/3”, 1/2”, etc.) it is important for the lens MTF to be considered for the appropriate imaging chip.

In the graph shown in our example the vertical axis represents normalised resolution, relative to the spatial resolution the imaging chip has. This can also be expressed in pixel *pairs per millimetre* (pp/mm). Basically the number of pixels per width or height an imaging chip has is divided by the width or height in mm and the result then divided by two. So for example, a 1/3” chip, with dimensions 4.8x3.6mm that is used for PAL standard definition signal would typically be 752 x 582 active pixels. If we take the vertical pixels (582) and divide them by 3.6mm we get nearly 162 pixels per mm. Divided this by 2, gives us 81 lp/mm, which is the maximum resolution that this chip can produce with a matching lens. Clearly, if we have a lens with lower lp/mm, then, this would be the first bottleneck in the systems resolution.



The vertical axis in the graphical example shown here is the normalised resolution relative to this chip.
The horizontal axis reprsents the distance from the centre of the imaging chip, hence 10 would mean 10mm from the centre of the optical axis, e.g. of the centre of the imaging chip. The example shown here is taken from the photographic industry where chips sizes are much larger (typically 23.4x15.6mm) hence the horizontal axis shows more mm.

This might be a very important detail to note in CCTV, as we have quite a large variety of imaging chips, starting from 1/4” up to 2/3” with the megapixel cameras. Each different imaging chip size might require different lens made to suit such chip size.

Getting back to the ICU units, in CCTV we are interested to quantify ICU based mostly on face identification, so it is natural to expect that such a face would not appear in the middle of the imaging chip (where the MTF is the highest) but rather it would go near the top of the imaging chip where a person head is expecte to be projected if a full person's height is projected on th eimaging chip.
This is why we suggest to use “realistic” MTF values at the point which is approximately 2/3 of the imaging chip half height (from the centre up), approximately coinciding with the area where a person face would occur.
For example, if we use a 1/3” chip, which is known to have dimensions of 4.8x3.6mm, than half of the imaging chip height is 1.8mm, and if we would to use MTF characteristics of a lens for CCTV evaluation of ICU units, we would read the MTF value at around the 1.2mm (2/3 of 1.8mm) from the centre of the imaging chip.

Once the MTF of a particular lens at the specific focal length setting (vari-focal length lenses would have various MPTF characteristic at various focal length setting) is known, represented with a normalised Modulation Transfer Function (MTF), we recommend to use the % of the MTF at the 2/3 from the centre of the imaging chip.
For illustration purposes, if the graph example shown above would refer to a camera with imaging chip height of, for example, 15mm, than 2/3 from the centre of the chip would be 5mm (7.5mm/3=2.5x2=5mm). At such a point the MTF of the blue thick dotted line (let's assume this the focal length we require) shows approximately 0.9 of the MTF. So that the ICU value would be reduced by this factor. So, if we had an ICU value with all the previousluy discussed factors, for example, equal to 0.9 ICU, the resulting ICU which includes lens quality factor would be now 0.81 ICU.

If the MTF factor is not included in the ICU unit of a particular camera/system (simply because it is not available or not known for the particular lens) we must then include in the ICU statement that lens parameters are excluded from such and such ICU value. Since lens ICU factorisation will always be equal or less than 1, for practical purposes it is suggested that if unknown lens is used, an assumtpion of ICU of 0.8 is used, unless proven otherwise.

6 Introducing the latency parameter (ICUl)

This is a parameter that could potentially be related to the 200ms again.

The latency is a combination of sampling/encoding, decoding and network latency.

The most important inclusion of latency ICU factor would be when designing IP systems to work in a LAN, and when there are PTZ to control there. Operators need immediate feedback of where the camera is being driven to, but the latency of encoding, decoding and the network might influence the response time and system efficiency.

Similarly, if fixed cameras are used, if there is “live” monitoring of megapixel cameras for example, we would like to know what is the delay between the real events in the hotel foyer, for example, and the image an operator is receiving on his/her screen.

There is no doubt, any digital system will introduce latency, the question is what is allowable and in what kind of application.

Present experience with large digital CCTV with thousands of cameras and recorders shows that when the latency can be brought within 200ms or below, and in such cases operators do not notice any degradation of their control. They can successfully use the system (follow incidents and people). Anything above this number might be causing some issues, but again, this is only noticeable when being interactive with the “live” images such as PTZ control. There is almost no issue with latency of even a couple of seconds if seen on the “live” vision screen for “non-interactive” system.

The following factors are suggested by Standard Australia:

Latency of 100ms or less	ICU = 1 (100%).
Latency between 100ms and 200ms	ICU = 0.9 (90%).
Latencies between 200ms and 400ms	ICU = 0.8 (80%).
Latencies between 400ms and 1 second	ICU = 0.7 (70%).
Latencies between 1 second and 3 seconds	ICU = 0.6 (60%).
And finally latencies over 3 seconds	ICU = 0.5 (50%).

The intended factorisation for latencies longer than 3 should be “capped” at 0.5, with explanation that it is better to have an image even if delayed over 5 seconds (which could be the case with large megapixels cameras) rather than having no image at all.

This last ICU factorisation which includes latency in practice would only be required for systems where live interaction is needed or desired.
If such a requirement is not needed for a system, this should clearly be stated in a tender documents, so that this factor doesn't deteriorate the total system's ICU.

7 Putting it all together

Using the above definitions, any digital system can have it's quality measured or requested, where all important factors would be included. The importance of this approach is that it includes mechanisms to put all systems on levelled plain field, taking into account all CCTV important parameters.

The higher the ICU the better, but this would define the system cost, camera and lens used, as well as compression applied. This would refer to each and separate camera signal going through the system. For example, a tender might require a total 0.9 ICU (less the latency ICU) for cameras allocated for face identification, whereas 1.1 ICU (less the latency ICU) might be required for number plate recognition.
Another system, where PTZ control is intended via the network, might be specified to have an ICU of 1.8 for example, with intention to identify playing cards and chips on a casino gaming table.

It is envisaged that when a digital system quality in ICU units is expressed as a system measured quality, or requested to be included in a tender documentation, it is most important to have all factors included:

ICU = ICUr x ICUi x ICUc x ICUo x ICUl

where ICUr is derived from the angle of view and imaging chip pixel count, ICUi for images per second, ICUc for compression, ICUo for the optics, and if needed ICUl for the latency. Only the latency factor might be ommitted if the digital CCTV system doesn't have PTZ control or live monitoring, but even in this case, it is expected that a statement is made that ICUl is not included.

All CCTV manufacturers, suppliers and integrators are encouraged to obtain MTF characteristic for their lenses, and in case this is not available for any particular lens, it would be safely assumed that the MTF is 0.8 ICUo.

The same applies for obtaining a PSNR characteristic for the compression used by the particular encoder/compressor. In cases where such a characteristic is not available, this standards recommends that for safety an assumption is made that the ICUc is given a value of 0.8, like with the lenses.