UK £12.50

US $20.00

The Engineer’s Guide to Motion Compensation

by John Watkinson

HANDBOOK
SNELL & WILCOX
SERIES

The Engineer's Guide to Motion Compensation

by John Watkinson

John Watkinson is an independent author, journalist and consultant in the broadcast industry with more than 20 years of experience in research and development.

With a BSc (Hons) in Electronic Engineering and an MSc in Sound and Vibration, he has held teaching posts at a senior level with The Digital Equipment Corporation, Sony Broadcast and Ampex Ltd., before forming his own consultancy.

Regularly delivering technical papers at conferences including AES, SMPTE, IEE, ITS and Montreux, John Watkinson has also written numerous publications including "The Art of Digital Video", "The Art of Digital Audio" and "The Digital Video Tape Recorder."

SNELL & WILCOX
Engineering with Vision

INTRODUCTION

There are now quite a few motion compensated products on the market, yet they do not all work in the same way. The purpose of this document is to clarify the confusion surrounding motion estimation by explaining clearly how it works, both in theory and in practice.

Video from different sources may demonstrate widely varying motion characteristics. When motion portrayal is poor, all types of motion compensated devices may appear to perform similarly, despite being widely divergent in approach. This booklet will explain motion characteristics in such a way as to enable the reader select critical material to reveal tangible performance differences between motion compensation systems.

Motion estimation is a complex subject which is ordinarily discussed in mathematical language. This is not appropriate here, and the new explanations which follow will use plain English to make the subject accessible to a wide range of readers. In particular a new non-mathematical explanation of the Fourier transform has been developed which is fundamental to demystifying phase correlation.

CONTENTS

Section 1 – Motion in Television **Page 2**

1.1 Motion and the eye
1.2 Motion in video systems
1.3 Conventional standards conversion
1.4 Motion compensated standards conversion
1.5 Methods of motion estimation
1.5.1 Block matching
1.5.2 Gradient methods
1.5.3 Phase correlation

Section 2 – Motion estimation using phase correlation **Page 23**

2.1 Phase correlation
2.2 Pre-processing
2.3 Motion estimation
2.4 Image correlation
2.5 Vector assignment
2.6 Obscured and revealed backgrounds

Section 3 – The Alchemist **Page 39**

3.1 Standards conversion
3.2 Motion compensation
3.3 Interpolation

Section 4 – Further applications **Page 49**

4.1 Gazelle - flow motion system
4.2 Noise reduction
4.3 Oversampling displays
4.4 Telecine Transfer

SECTION 1 - MOTION IN TELEVISION

1.1 Motion and the eye

The human eye is not unlike a CCD camera in that the retina is covered with a large number of discrete sensors which are used to build up an image. The spacing between the sensors has a great bearing on the resolution of the eye, in the same way as the number of sensors in a CCD chip. However, the eye acts to a degree as if it were A.C. coupled so that its response to low spatial frequencies (the average brightness of a scene) falls.

The eye also has a temporal response taking the form of a lag known as persistence of vision. The effect of the lag is that resolution is lost in areas where the image is moving rapidly over the retina

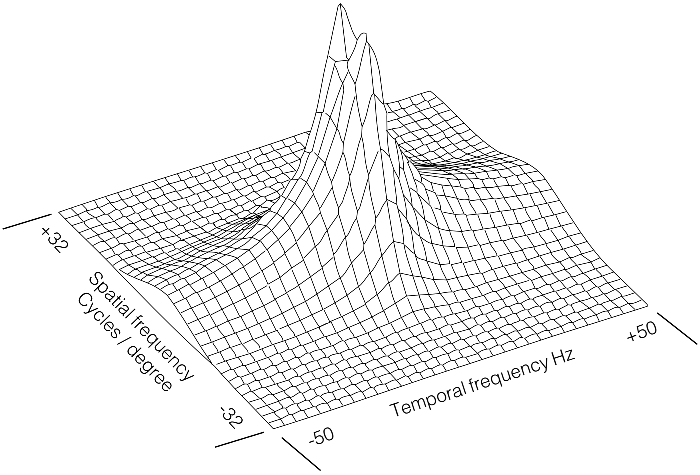


Fig 1.1.1 The response of the eye shown with respect to temporal and spatial frequencies. Note that even slow relative movement causes a serious loss of resolution. The eye tracks moving objects to prevent this loss.

Fig 1.1.1 shows the response of the eye in two dimensions; temporal frequency (field rates are measured in this) and spatial frequency (resolution). The response falls off quite rapidly with temporal frequency; a phenomenon known as persistence of vision or motion blur.

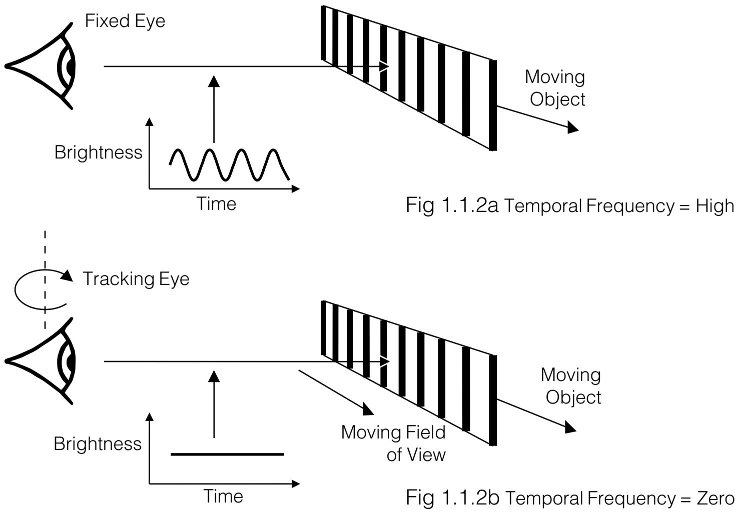


Fig 1.1.2a Temporal Frequency = High

Fig 1.1.2b Temporal Frequency = Zero

Fig 1.1.2 a A detailed object moves past a fixed eye, causing temporal frequencies beyond the response of the eye. This is the cause of motion blur.
b The eye tracks the motion and the temporal frequency becomes zero. Motion blur cannot then occur.

Thus a fixed eye has poor resolution of moving objects. Many years ago this was used as an argument that man would never be able to fly because he would not be able to see properly at speeds needed for flight. Clearly the argument is false, because the eye can move to follow objects of interest. Fig 1.1.2 shows the difference this makes. At a) a detailed object moves past a fixed eye. It does not have to move very fast before the temporal frequency at a fixed point on the retina rises beyond the temporal response of the eye. This is motion blur. At b) the eye is following the moving object and as a result the temporal frequency at a fixed point on the retina is zero; the full resolution is then available because the image is stationary with respect to the eye. In real life we can see moving objects in some detail unless they move faster than the eye can follow. Exceptions are in the case of irregular motion which the eye cannot track or rotational motion; the eyeball cannot rotate on the optical axis!

1.2 Motion in video systems

Television viewing is affected by two distinct forms of motion. One of these is the motion of the original scene with respect to the camera and the other is the motion of the eye with respect to the display. The situation is further complicated by the fact that television pictures are not continuous, but are sampled at the field rate.

According to sampling theory, a sampling system cannot properly convey frequencies beyond half the sampling rate. If the sampling rate is considered to be the field rate, then no temporal frequency of more than 25 or 30 Hz can be handled (12 Hz for film). With a stationary camera and scene, temporal frequencies can only result from the brightness of lighting changing, but this is unlikely to cause a problem. However, when there is relative movement between camera and scene, detailed areas develop high temporal frequencies, just as was shown in Fig 1.1.2 for the eye. This is because relative motion results in a given point on the camera sensor effectively scanning across the scene. The temporal frequencies generated are beyond the limit set by sampling theory, and aliasing should be visible.

However, when the resultant pictures are viewed by a human eye, this aliasing is not perceived because, once more, the eye attempts to follow the motion of the scene.

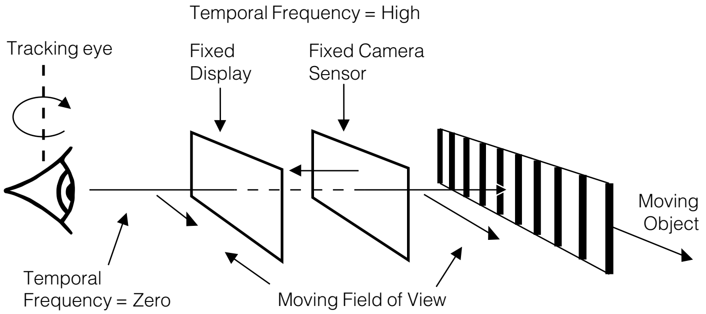


Fig 1.2.1 An object moves past a camera, and is tracked on a monitor by the eye. The high temporal frequencies cause aliasing in the TV signal, but these are not perceived by the tracking eye as this reduces the temporal frequency to zero. Compare with Fig 1.1.2

Fig 1.2.1 shows what happens when the eye follows correctly. Effectively the original scene and the retina are stationary with respect to one another, but the camera sensor and display are both moving through the field of view. As a result the temporal frequency at the eye due to the object being followed is brought to zero and no aliasing is perceived by the viewer due to the field rate sampling.

Whilst this result is highly desirable, we have not actually circumvented sampling theory, because the effect only works if several assumptions are made, including the requirement for the motion to be smooth.

What is seen is not quite the same as if the scene were being viewed through a piece of moving glass because of the movement of the image relative to the camera sensor and the display. Here, temporal frequencies do exist, and the temporal aperture effect (lag) of both will reduce perceived resolution. This is the reason that shutters are sometimes fitted to CCD cameras used at sporting events. The mechanically rotating shutter allows light onto the CCD sensor for only part of the field period thereby reducing the temporal aperture. The result is obvious from conventional photography in which one naturally uses a short exposure for moving subjects. The shuttered CCD camera effectively has an exposure control. On the other hand a tube camera displays considerable lag and will not perform as well under these circumstances.

It is important to appreciate the effect of camera type and temporal aperture as it has a great bearing on how to go about assessing or comparing the performance of motion compensated standards converters. One of the strengths of motion compensation is that it improves the resolution of moving objects. This improvement will not be apparent if the source material being used has poor motion qualities in the first place. Using unsuitable (i.e. uncritical) material could result in two different converters giving the same apparent performance, whereas more critical material would allow the better convertor to show its paces.

As the greatest problem in standards conversion is the handling of the time axis, it is intended to contrast the time properties of various types of video.

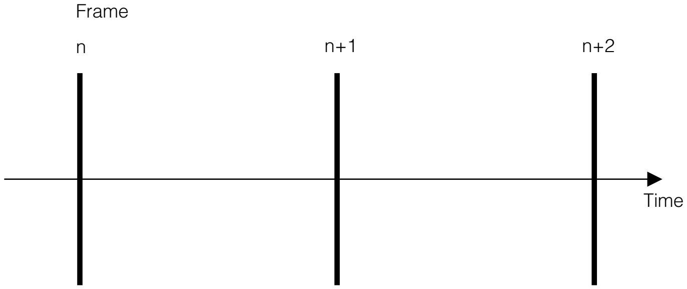


Fig 1.2.2 The spatio-temporal characteristic of film. Note that each frame is repeated twice on projection.

Fig 1.2.2 shows the simplest time axis, that of film, where entire frames are simultaneously exposed, or sampled, and the result is that the image is effectively at right angles to the time axis. When displayed, each frame of a film is generally projected twice. The result with a moving object is that the motion is not properly portrayed and there is judder

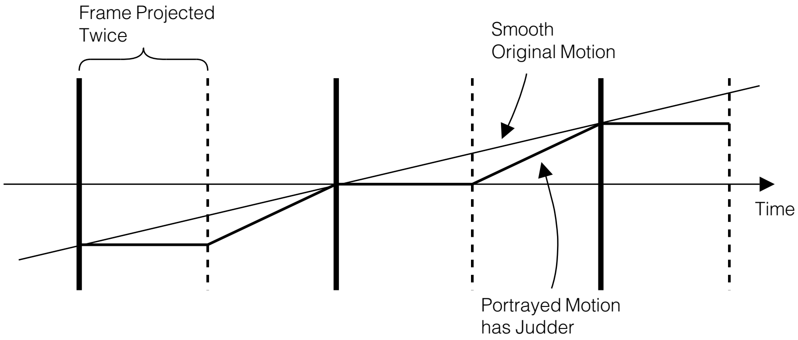


Fig 1.2.3 The frame repeating results in motion judder as shown here.

Fig 1.2.3 shows the origin of the judder. It should be appreciated that the judder is not present on the film, but results from the projection method. Information from one place on the time axis appears in two places. Tube cameras do not sample the image all at once, but scan it from top to bottom, integrating it over the field period. Scanning may be progressive or interlaced.

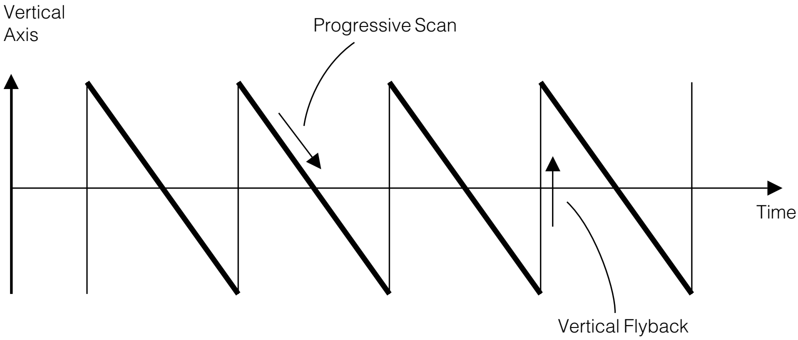


Fig 1.2.4 The spatio temporal characteristic of progressive scan TV. Note the characteristic tilt of the image planes.

Fig 1.2.4 shows the time axis of progressive scan television cameras and CRT displays. The vertical scan takes a substantial part of the frame period and so the image is tilted with respect to the time axis. As both camera and display have the same tilt, the tilt has no effect on motion portrayal.

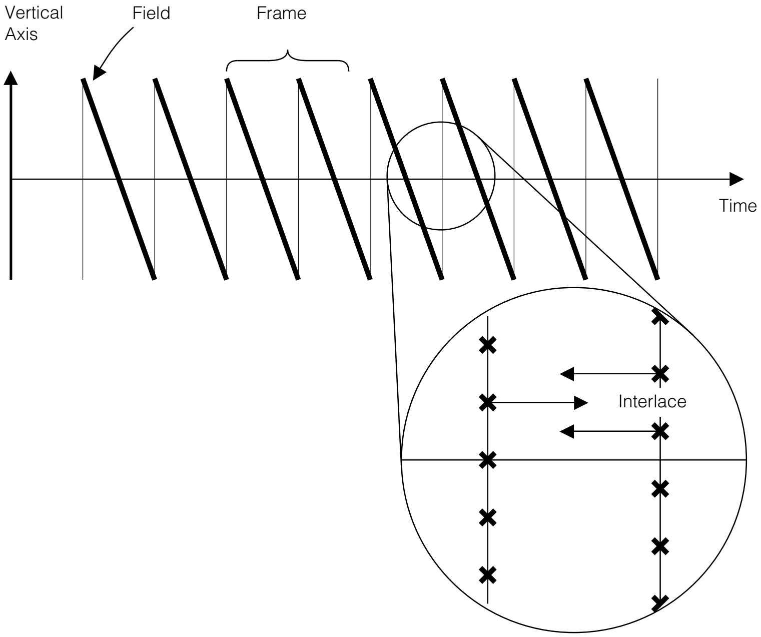


Fig 1.2.5 The result of interlaced scan.

Fig 1.2.5 shows the time axis of interlaced scan cameras and displays. The vertical scan is twice as fast and so the tilt is now over a field period. Again as camera and CRT display have the same tilt, the tilt does not affect motion portrayal. In CCD cameras, the image is sampled at once, like film, and thus the fields are at right angles to the time axis. If displayed on a scanning CRT, the time axis is distorted such that objects are displayed later than they should be towards the bottom of the screen. Transversely moving objects have speed dependent sloping verticals, as shown in Fig 1.2.6.

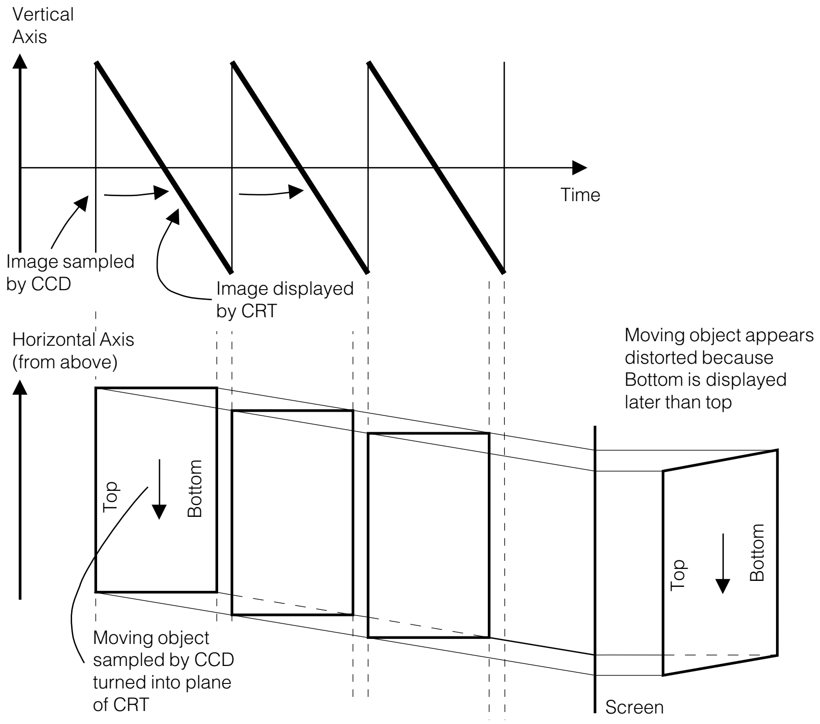

Vertical Axis
Time
Image sampled by CCD
Image displayed by CRT
Horizontal Axis (from above)
Top
Bottom
Moving object appears distorted because Bottom is displayed later than top
Top
Bottom
Moving object sampled by CCD turned into plane of CRT
Screen

Fig 1.2.6 Displaying sampled CCD images on a scanned CRT results in distortion in the presence of motion.

The same effect is evident if film is displayed on a CRT via telecine. The film frames are sampled, but the CRT is scanned. In telecine each frame results in two fields in 50 Hz standards and this will result in judder as well. In 60Hz telecine odd frames result in two fields, even frames result in three fields; the well known 3:2 pulldown. Motion portrayal (or lack of it) in this case is shown in Fig 1.2.7. Thus CCD cameras to some extent and telecines to a great extent are not compatible with scanning (CRT) displays.

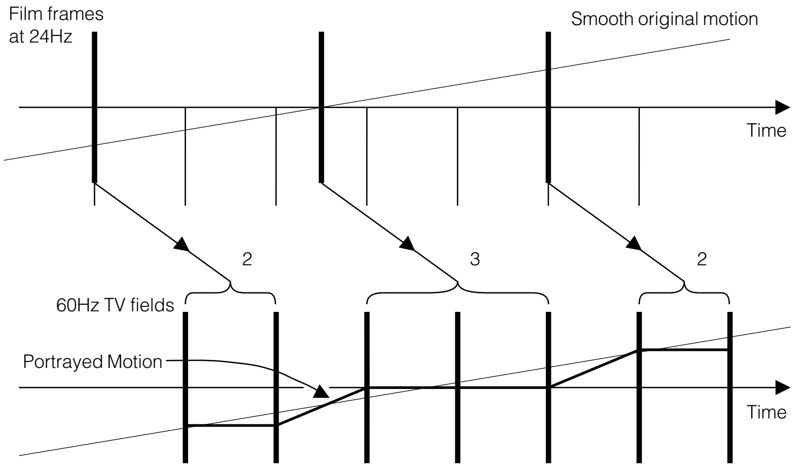


Fig 1.2.7 Telecine machines must use 3:2 pulldown to produce 60 Hz field rate video

1.3 Conventional standards conversion

Standards conversion is primarily concerned with changing the number of lines in the field and the field rate. The number of lines in the field can readily be changed using interpolation and this is not a difficult process, as will be seen. More to the point, it can be performed with a minimum of degradation provided an appropriate degree of complexity is adopted in the computation. This is because the number of lines in the field, i.e. the vertical sampling frequency, compares much more favourably with the vertical frequencies likely to be in the image, and there will be much less aliasing than in the time axis.

In contrast, the field rate change is extremely difficult not least because the field rate is so low compared to the temporal frequencies which are created when objects move. Thus temporal aliasing in the input signal is the norm rather than the exception. It was seen in section 1.1 that the eye would not see this aliasing in a normal television system because it would follow motion. However, a conventional standards convertor is not transparent to motion portrayal, and the effect is judder and loss of resolution in the converted picture.

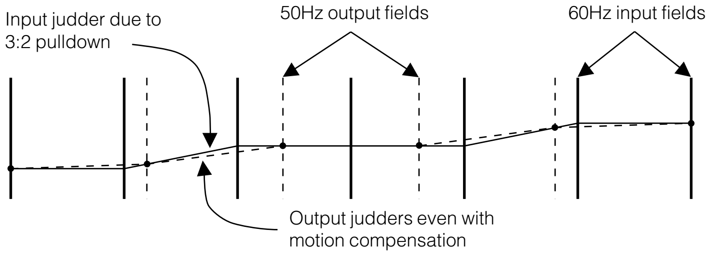


Fig 1.2.8 A motion compensated standards converter cannot remove judder which is present on the input signal. Here a 60Hz input signal from a 3:2 pulldown telecine machine contains irregular motion (which was not on the film) and the 50Hz output will also display judder.

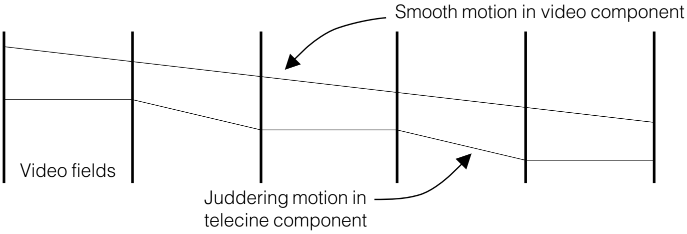


Fig 1.2.9 If material from telecine and video sources is overlayed, the video material appears judder free, but the telecine component judders.

It should be appreciated that a motion compensated standards convertor is designed to eliminate judder from the conversion process. As a result, motion on the input is faithfully reproduced at the output. Unfortunately video from conventional telecine machines actually contains juddering motion, as was seen in Fig 1.2.7. A motion compensated standards convertor cannot distinguish between telecine judder and actual source picture motion. Externally generated judder appears at the output as shown in Fig 1.2.8. This is an example of the well known "garbage in – garbage out" phenomenon. In some post production processes, video from cameras is cut into video from telecine, or vice versa. The result is that the two components of the combined image have different motion, as shown in Fig 1.2.9. The video component is judder free, but the telecine component judders. With such an input, a motion compensated standards convertor passes the telecine judder and the result is bound to be disappointing. If standards conversion of such material is contemplated, there is only one way to obtain good results, and this is to employ motion compensation at the telecine machine as is described in section 4.4.

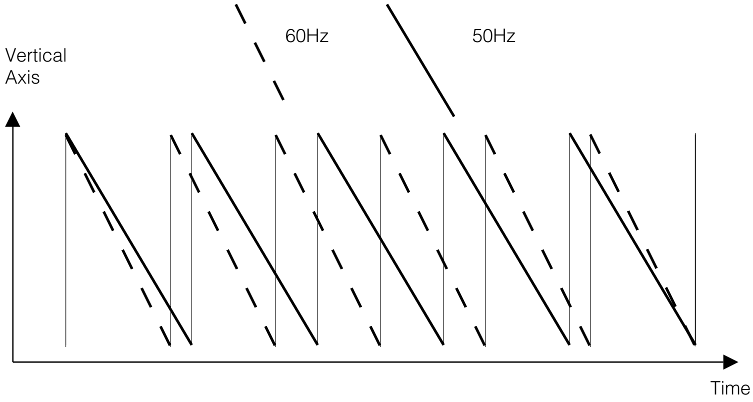


Fig 1.3.1 The standards conversion problem is simply to transfer images from one set of planes in the spatio-temporal volume to another.

Fig 1.3.1 shows what happens on the time axis in a conversion between 60 Hz and 50 Hz (in either direction). Fields in the two standards appear in different planes cutting through the spatio-temporal volume, and the job of the standards convertor is to interpolate between input planes in one standard in order to estimate what an intermediate plane in the other standard would look like. With still images, this is easy, because planes can be slid up and down the time axis with no ill effect. 60 fields a second can be turned into 50 fields a second simply by dumping one field in six. However, if the program material contains motion, this approach fails. Fig 1.3.2 shows that field dumping results in a 10 Hz jerkiness in the movement.

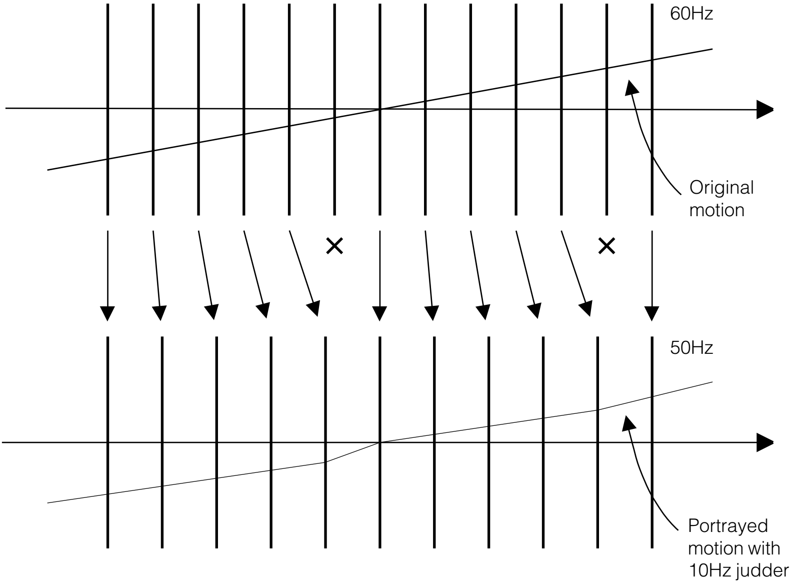
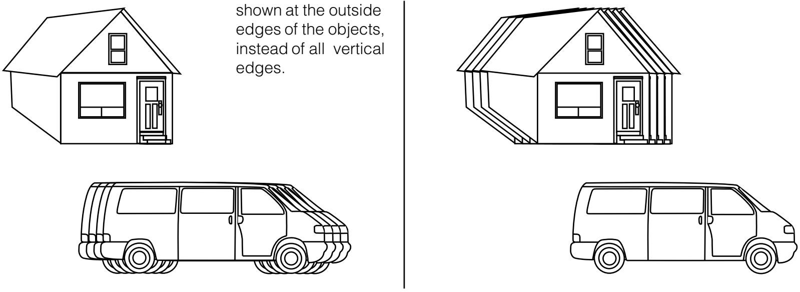


Fig 1.3.2 Simple field dumping produces an unacceptable 10 Hz component and jerky motion portrayal.

The 10 Hz component is a beat or alias frequency between the two field rates and practical converters attempt to filter it out by incorporating low pass filtering in the time axis. The frequency response of the filter must extend no further than permitted by sampling theory. In other words the temporal frequency response must not exceed one half of the lowest sampling rate, i.e. 25 Hz. Filtering to this low frequency can only be achieved by having an impulse response which spreads over at least four fields along the time axis, hence the conventional four-field standards convertor. Four field converters interpolate through time, using pixel data from four input fields in order to compute the likely values of pixels in an intermediate output field. This eliminates the 10 Hz effects, but does not handle motion transparently.



In the interests of clarity, judder is only shown at the outside edges of the objects, instead of all vertical edges.

a) Fixed camera

b) Panning camera

Fig 1.3.3 a Conventional four field convertor with moving object produces multiple images.

b If the camera is panned on the moving object, the judder moves to the background.

Fig 1.3.3a) shows that if an object is moving, it will be in a different place in successive fields. Interpolating between several fields results in multiple images of the object. The position of the dominant image will not move smoothly, an effect which is perceived as judder. If the camera is panning the moving object, it will be in much the same place in successive fields and Fig 1.3.3b) shows that it will be the background which judders.

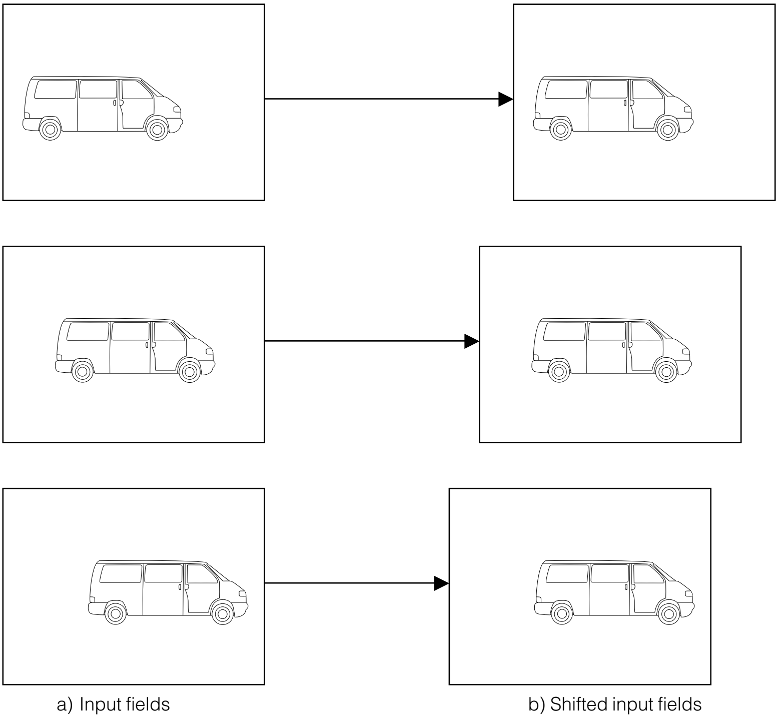


Fig 1.4.1 a Successive fields with moving object.
b Motion compensation shifts the fields to align position of the moving object.

1.4 Motion compensated standards conversion

The basic principle of motion compensation is quite simple. In the case of a moving object, it appears in different places in successive source fields. Motion compensation computes where the object will be in an intermediate target field and then shifts the object to that position in each of the source fields. Fig 1.4.1a) shows the original fields, and Fig 1.4.1b) shows the result after shifting. This explanation is only suitable for illustrating the processing of a single motion such as a pan.

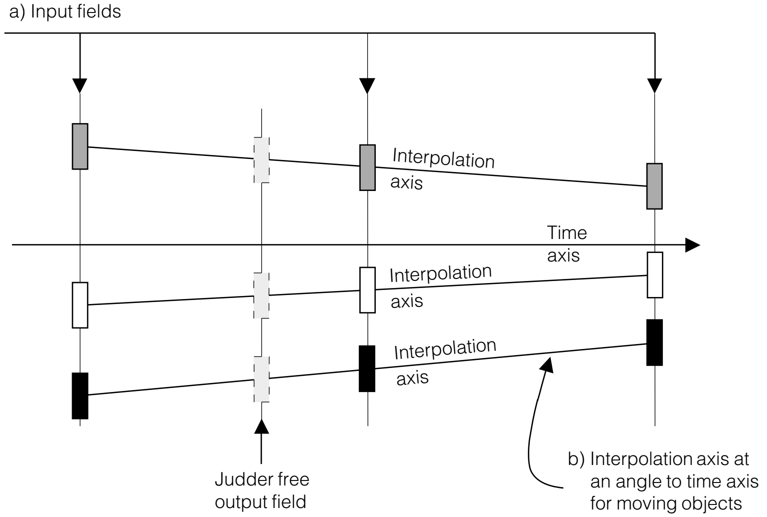


Fig 1.4.2 a Input fields with moving objects.
b Moving the interpolation axes to make them parallel to the trajectory of each object.

An alternative way of looking at motion compensation is to consider what happens in the spatio-temporal volume. A conventional standards convertor interpolates only along the time axis, whereas a motion compensated standards convertor can swivel its interpolation axis off the time axis. Fig 1.4.2a) shows the input fields in which three objects are moving in a different way. At b) it will be seen that the interpolation axis is aligned with the trajectory of each moving object in turn.

This has a dramatic effect. Each object is no longer moving with respect to its own interpolation axis, and so on that axis it no longer generates temporal frequencies due to motion and temporal aliasing cannot occur. Interpolation along the correct axes will then result in a sequence of output fields in which motion is properly portrayed. The process requires a standards convertor which contains filters which are modified to allow the interpolation axis to move dynamically within each output field. The signals which move the interpolation axis are known as motion vectors. It is the job of the motion estimation system to provide these motion vectors. The overall performance of the convertor is determined primarily by the accuracy of the motion vectors. An incorrect vector will result in unrelated pixels from several fields being superimposed and the result will be unsatisfactory.

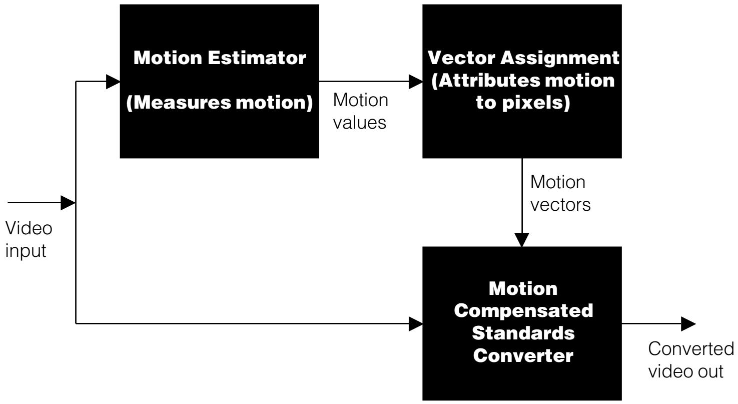


Fig 1.4.3 The essential stages of a motion compensated standards convertor.

Fig 1.4.3 shows the sequence of events in a motion compensated standards convertor. The motion estimator measures movements between successive fields. These motions must then be attributed to objects by creating boundaries around sets of pixels having the same motion. The result of this process is a set of motion vectors, hence the term vector assignment. The motion vectors are then input to a specially designed four field standards convertor in order to deflect the inter-field interpolation axis. Note that motion estimation and motion compensation are two different processes.

1.5 Methods of motion estimation

In any motion compensated device, one necessary process is motion estimation, in which successive source images are analysed to identify objects which are in motion. The motion of an object is described by a motion vector; a two dimensional parameter having length and direction. There must also be a process called vector assignment, in which the appropriate motion vectors are assigned to every pixel location in the picture to make a vector field. The third stage is the actual motion compensation process which uses the vector fields to direct the axis of the filtering to be parallel to the optic flow axis.

Whilst none of these processes are trivial, the motion compensation stage is fairly straightforward because the manipulations taking place are basically what happens in a DVE. In motion estimation it is not the details of the problem which cause difficulty, but the magnitude. All of the available techniques need a fair amount of processing power. Here the criterion is to select an algorithm which gives the best efficiency at the desired performance level. In practice the most difficult section of a machine to engineer is the logic which interprets the output of the motion estimator. Another area in which care is needed is in the pre-processing which is necessary to allow motion estimation to take place between fields. The use of interlace means that one field cannot be compared directly with the next because the available pixels are in different places in the image.

As television signals are effectively undersampled in the time axis, it is impossible to make a motion estimator which always works. Difficult inputs having periodic structures or highly irregular motion may result in incorrect estimation. Since any motion estimation system will fail, it is important to consider how such a failure is handled. It is better to have a graceful transition to conventional conversion than a spectacular failure. Motion estimation techniques can also be compared on the allowable speed range and precision. Standards conversion of sporting events will require a large motion range, whereas correction of film weave will require a range of a couple of pixels at worst, but will need higher sub-pixel accuracy. A slow-motion system needs both. Finally estimators can be compared on the number of different motions which can be handled at once. There are a number of estimation techniques which can be used, and these are compared below. It is dangerous to generalise about actual products as a result of these descriptions. The terminology is not always used consistently, techniques are seldom used alone or in a pure form, and the quality of the vector assignment process is just as important, perhaps more important, in determining performance.

1.5.1 Block matching

Fig 1.5.1 shows a block of pixels in an image. This block is compared, a pixel at a time, with a similarly sized block b) in the same place in the next image. If there is no motion between fields, there will be high correlation between the pixel values. However, in the case of motion, the same, or similar pixel values will be elsewhere and it will be necessary to search for them by moving the search block to all possible locations in the search area. The location which gives the best correlation is assumed to be the new location of a moving object.

Whilst simple in concept, block matching requires an enormous amount of computation because every possible motion must be tested over the assumed range. Thus if the object is assumed to have moved over a sixteen pixel range, then it will be necessary to test 16 different horizontal displacements in each of sixteen vertical positions; in excess of 65,000 positions. At each position every pixel in the block must be compared with every pixel in the second picture.

In typical video, displacements of twice the figure quoted here may be found, particularly in sporting events, and the computation then required becomes enormous.

One way of reducing the amount of computation is to perform the matching in stages where the first stage is inaccurate but covers a large motion range but the last stage is accurate but covers a small range. The first matching stage is performed on a heavily filtered and subsampled picture, which contains far fewer pixels. When a match is found, the displacement is used as a basis for a second stage which is performed with a less heavily filtered picture. Eventually the last stage takes place to any desired accuracy. This hierarchical approach does reduce the computation required, but it suffers from the problem that the filtering of the first stage may make small objects disappear and they can never be found by subsequent stages if they are moving with respect to their background. Many televised sports events contain small, fast moving objects. As the matching process depends upon finding similar luminance values, this can be confused by objects moving into shade or by fades. As block matching relies on comparison of pixel values, the motion vectors it produces can only be accurate to the nearest pixel. This is a disadvantage for standards conversion, but is an advantage for data reduction applications where block matching is frequently employed. Inaccuracies in motion estimation are not important in data reduction because they are inside the error loop and are cancelled by sending appropriate frame difference data. If a small moving object is missed by the motion estimator, the only effect is that the difference data increase slightly.

Section 1.5.2 Hierarchical Spatial Correlation

This technique is a form of block matching, but was given the name of spatial correlation to distinguish it from the generic technique described above. Considering the example of Fig 1.5.1, there is a conflict over the best search block size. A small block allows many different motions to be measured, but has a high probability of achieving good correlation by chance, resulting in a false motion measurement. On the other hand, a large block avoids false matches, but fewer motions can be handled. A hierarchical system can be used to resolve the conflict. Fig 1.5.2 shows that the initial block size is large, but in the next stage the block is subdivided and the motion from the first stage is used as a basis for four individual motion estimates. Subsequent stages can divide the search block size down to any desired accuracy. The technique can be combined with motion prediction from previous fields.

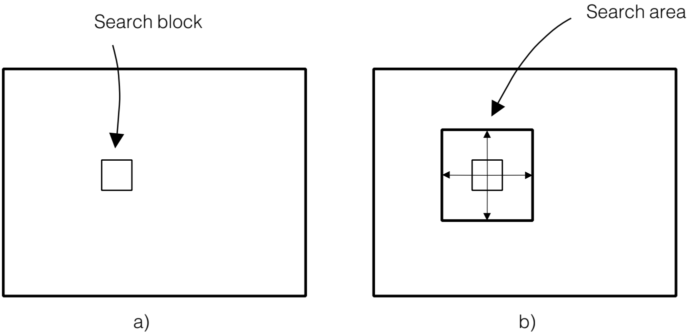


Fig 1.5.1 a The search block is a reference and a similar block is expected in the next picture.
b The search block is moved around the search area in order to find the best match which is assumed to correspond to the motion between pictures

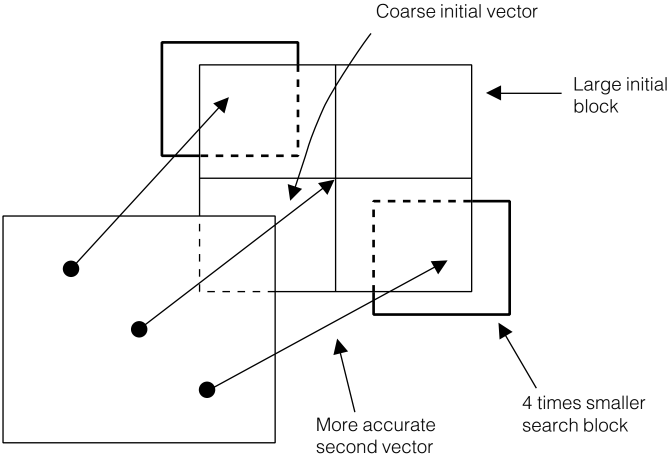


Fig 1.5.2 In hierarchical spatial correlation, an initial block match is made with a large block which is then divided into four. The coarse initial vector is used as a basis for four separate more accurate searches. As many iterations as needed may be used.

1.5.3 Gradient methods

At some point in an image, the function of brightness with respect to distance across the screen will have a particular slope, known as a spatial luminance gradient. It is possible to estimate the motion between images by locating a similar gradient in the next image. Alternatively, the spatial gradient may be compared with the temporal gradient as shown in Fig 1.5.3. If the associated picture area is moving, the slope will traverse a fixed point on the screen and the result will be that the brightness now changes with respect to time. For a given spatial gradient, the temporal gradient becomes steeper as the speed of movement increases. Thus motion speed can be estimated from the ratio of the spatial and temporal gradients.

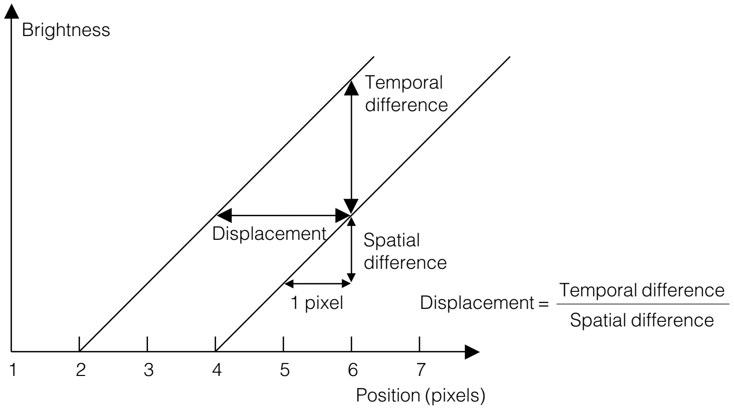


Fig 1.5.3 Gradient Method measurement technique

In practice there are numerous difficulties. A similar spatial gradient may be found in the next image which is due to a different object, and a false match results. When an object moves in a direction parallel to its side, there is no temporal gradient. When an object moves so as to obscure or reveal the background, the spatial gradient will change from field to field even if the motion is constant. Variations in illumination, such as when an object moves into shade, also cause difficulty. One of the greatest difficulties is accurate measurement of temporal gradients in the presence of temporal aliasing. If temporal aliasing due to moving objects were absent, gradient measurement would be easy, but motion compensation would not be necessary. Stated differently, the reason why motion compensation is necessary is the same reason that temporal gradient measurement is inaccurate. The accuracy of both gradient methods can be improved by using recursion. In this case the original estimate of the velocity is used as a basis for successive calculations. In highly detailed images, the motion range which can be handled between fields becomes very small and it becomes essential to use prediction of the motion from earlier fields. This causes difficulty at cuts. In periodic structures such as grilles, there are many places where the gradient is the same and handling of such images is poor.

1.5.4 Phase correlation

Phase correlation works by performing a spectral analysis on two successive fields and then subtracting all of the phases of the spectral components. The phase differences are then subject to a reverse transform which directly reveals peaks whose positions correspond to motions between the fields. The nature of the transform domain means that if the distance and direction of the motion is measured accurately, the area of the screen in which it took place is not. Thus in practical systems the phase correlation stage is followed by a matching stage not dissimilar to the block matching process. However, the matching process is steered by the motions from the phase correlation, and so there is no need to attempt to match at all possible motions. By attempting matching on measured motion only the overall process is made much more efficient.

One way of considering phase correlation is that by using the spectral analysis to break the picture into its constituent spatial frequencies the hierarchical structure of block matching at various resolutions is in fact performed in parallel. In this way small objects are not missed because they will generate high frequency components in the transform. A further advantage of phase correlation is that it is sub-pixel accurate. Although the matching process is simplified by adopting phase correlation, the spectral analysis requires complex calculations. The high performance of phase correlation would remain academic if it were too complex to put into practice. However, if realistic values are used for the motion speeds which can be handled, the computation required by block matching actually exceeds that required for phase correlation.

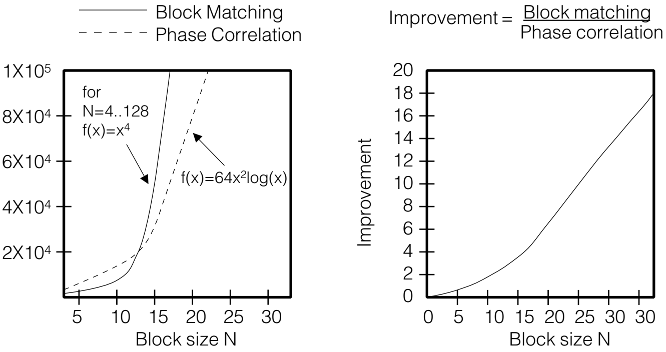


Fig 1.5.4 Relative computational complexity of phase correlation versus block matching.

The elimination of amplitude information from the phase correlation process ensures that motion estimation continues to work in the case of fades, objects moving into shade or flashguns firing.

Section 2 - MOTION ESTIMATION USING PHASE CORRELATION

Based on the conclusions of the previous section, Phase Correlation has a great deal to offer, but is a complex process, and warrants a detailed explanation.

The use of motion compensation enhances standards conversion by moving picture areas across the screen to properly portray motion in a different field structure (see fig 1.4.1). Motion compensation is the process of modifying the operation of a nearly conventional standards convertor. The motion compensation itself is controlled by parameters known as motion vectors which are produced by the motion estimation unit. It should be stressed that phase correlation is just one step, albeit a critical one, in the motion estimation process.

2.1 Phase correlation

The details of the Fourier transform are described in Appendix A. The use of this transform is vital to the concept of phase correlation. A one dimensional example will be given here by way of introduction. A line of luminance, which in the digital domain consists of a series of samples, is a function of brightness with respect to distance across the screen. The Fourier transform converts this function into a spectrum of spatial frequencies (units of cycles per picture width) and phases.

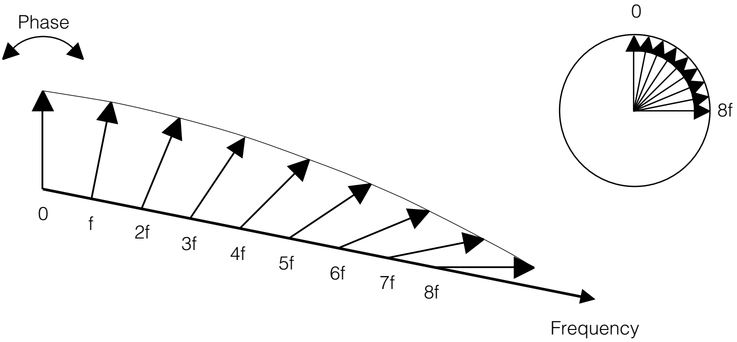


Fig 2.1.1 The definition of phase linearity is that phase shift is proportional to frequency. In linear phase systems the waveform is preserved, and simply moves in time or space.

All television signals must be handled in linear-phase systems. A linear phase system is one in which the delay experienced is the same for all frequencies. If video signals pass through a device which does not exhibit linear phase, the various frequency components of edges become displaced across the screen. Fig 2.1.1 shows what phase linearity means. If the left hand end of the frequency axis (0) is considered to be firmly anchored, but the right hand end can be rotated to represent a change of position across the screen, it will be seen that as the axis twists evenly the result is phase shift proportional to frequency. A system having this characteristic is said to have linear phase.

In the spatial domain, a phase shift corresponds to a physical movement. Fig 2.1.2 shows that if between fields a waveform moves along the line, the lowest frequency in the Fourier transform will suffer a given phase shift, twice that frequency will suffer twice that phase shift and so on. Thus it is potentially possible to measure movement between two successive fields if the phase differences between the Fourier spectra are analysed. This is the basis of phase correlation.

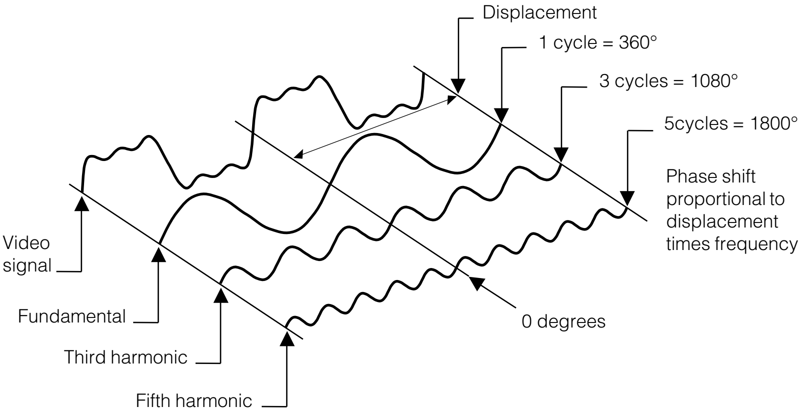


Fig 2.1.2 In a linear phase system, shifting the video waveform across the screen causes phase shifts in each component proportional to frequency.

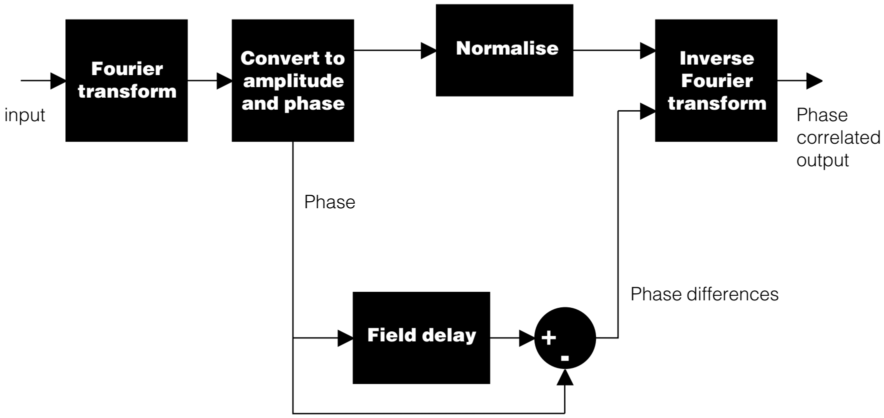


Fig 2.1.3 The basic components of a phase correlator.

Fig 2.1.3 shows how a one dimensional phase correlator works. The Fourier transforms of two lines from successive fields are computed and expressed in polar (amplitude and phase) notation. The phases of one transform are all subtracted from the phases of the same frequencies in the other transform. Any frequency component having significant amplitude is then normalised, i.e., boosted to full amplitude.

The result is a set of frequency components which all have the same amplitude, but have phases corresponding to the difference between two fields. These coefficients form the input to an inverse transform.

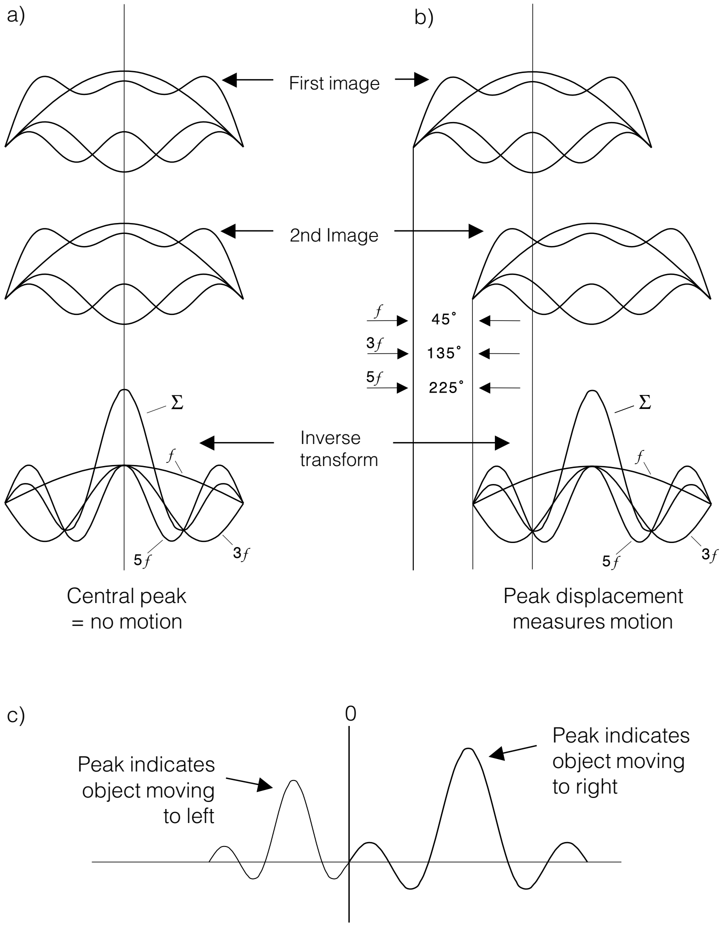


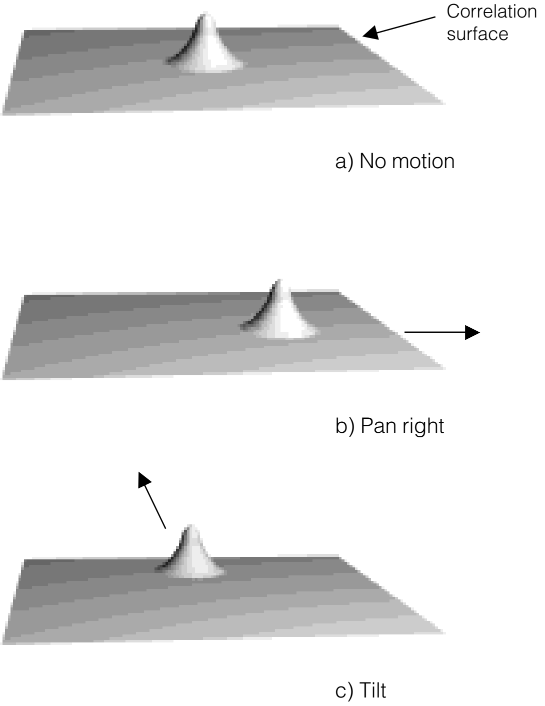
Fig 2.1.4 a The peak in the inverse transform is central for no motion.
b In the case of motion, the peak shifts by the distance moved.
c If there are several motions, each one results in a peak.

Fig 2.1.4a) shows what happens. If the two fields are the same, there are no phase differences between the two, and so all of the frequency components are added with zero degrees phase to produce a single peak in the centre of the inverse transform. If, however there was motion between the two fields, such as a pan, all of the components will have phase differences, and this results in a peak shown in

Fig 2.1.4b) which is displaced from the centre of the inverse transform by the distance moved. Phase correlation thus actually measures the movement between fields, rather than inferring it from luminance matches.

In the case where the line of video in question intersects objects moving at different speeds, Fig 2.1.4c) shows that the inverse transform would contain one peak corresponding to the distance moved by each object.

Whilst this explanation has used one dimension for simplicity, in practice the entire process is two dimensional. A two dimensional Fourier transform of each field is computed, the phases are subtracted, and an inverse two dimensional transform is computed, the output of which is a flat plane out of which three dimensional peaks rise. This is known as a correlation surface.



a) No motion

b) Pan right

c) Tilt

Fig 2.1.5 a A two dimensional correlation surface has a central peak when there is no motion.
b In the case of a pan, the peak moves laterally.
c A camera tilt moves the peak at right angles to the pan.

Fig 2.1.5 shows some examples of a correlation surface. At a) there has been no motion between fields and so there is a single central peak.

At b) there has been a pan and the peak moves across the surface. At c) the camera has been depressed and the peak moves upwards.

Where more complex motions are involved, perhaps with several objects moving in different directions and / or at different speeds, one peak will appear in the correlation surface for each object.

It is a fundamental strength of phase correlation that it actually measures the direction and speed of moving objects rather than estimating, extrapolating or searching for them.

However it should be understood that accuracy in the transform domain is incompatible with accuracy in the spatial domain. Although phase correlation accurately measures motion speeds and directions, it cannot specify where in the picture these motions are taking place. It is necessary to look for them in a further matching process. The efficiency of this process is dramatically improved by the inputs from the phase correlation stage.

2.2 Pre-processing

The input to a motion estimator for standards conversion consists of interlaced fields. The lines of one field lie between those of the next, making comparisons between them difficult. A further problem is that vertical spatial aliasing may exist in the fields. Until recently, motion estimation was restricted to inter-frame measurement because of these problems. Inter-field motion measurement is preferable because the more often the motion can be measured the more accurate the motion portrayal will be. This is now possible if appropriate pre-processing is used. Pre-processing consists of a combined filtering and interpolation stage which simultaneously removes the effects of interlace and vertical aliasing. The same filtering response is applied to the horizontal axis to make the accuracy the same in both cases.

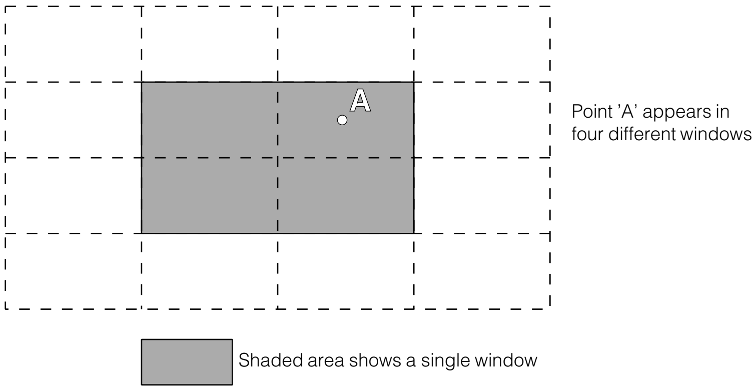

A
Point 'A' appears in four different windows
Shaded area shows a single window

Fig 2.2.1 The input fields are converted into overlapping windows. Each window is individually transformed.

The computation needed to perform a two dimensional Fourier transform increases dramatically with the size of the block employed, and so no attempt is made to transform the downsampled fields directly. Instead the fields are converted into overlapping blocks by the use of window functions as shown in Fig 2.2.1 The size of the window controls the motion speed which can be handled, and so a window size is chosen which allows motion to be detected up to the limit of human judder visibility. The window may be rectangular because horizontal motion is more common than vertical in real program material.

2.3 Motion estimation

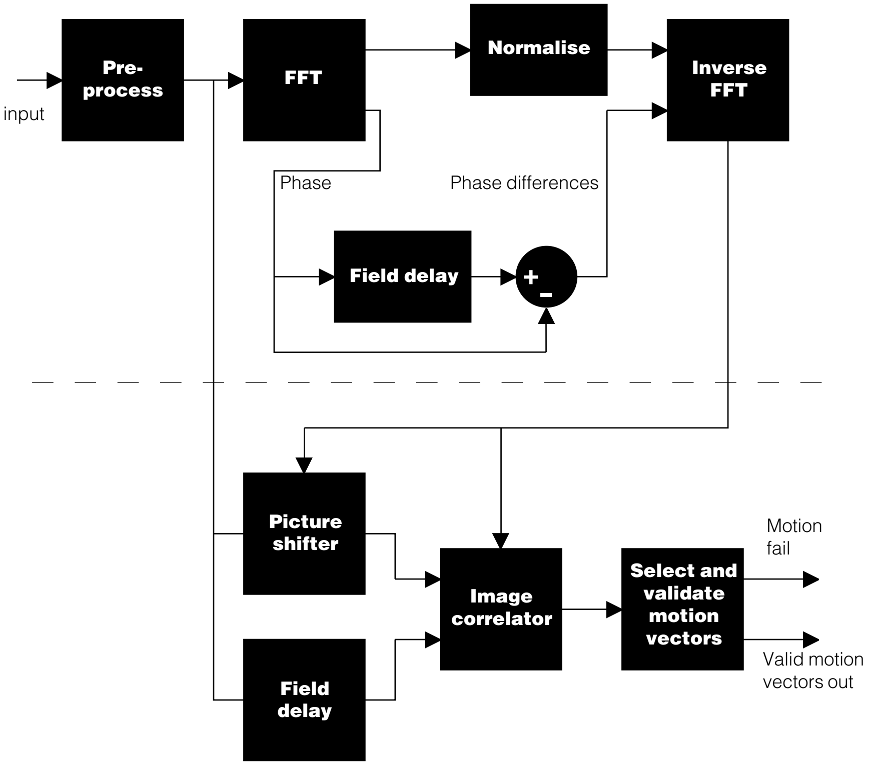


Fig 2.3.1 The block diagram of a phase correlated motion estimator

Fig 2.3.1 shows a block diagram of a phase correlated motion estimation system. Following the pre-processing, each windowed block is subject to a Fast Fourier Transform (FFT), and the output spectrum is converted to the amplitude and phase representation. The phases are subtracted from those of the previous field in each window, and the amplitudes are normalised to eliminate any variations in illumination or the effect of fades from the motion sensing. A reverse transform is performed, which results in a correlation surface. The correlation surface contains peaks whose positions actually measure distances and directions moved by some feature in the window. It is a characteristic of the Fourier transform that the more accurately the spectrum of a signal is known, the less accurately the spatial domain is known.

Thus the whereabouts within the window of the moving objects which gave rise to the correlation peaks is not known. Fig 2.3.2 illustrates the phenomenon. Two windowed blocks are shown in consecutive fields. Both contain the same objects, moving at the same speed, but from different starting points. The correlation surface will be the same in both cases. The phase correlation process therefore needs to be followed by a further process called vector assignment which identifies the picture areas in which the measured motion took place and establishes a level of confidence in the identification. This stage can also be seen in the block diagram of Fig 2.3.1

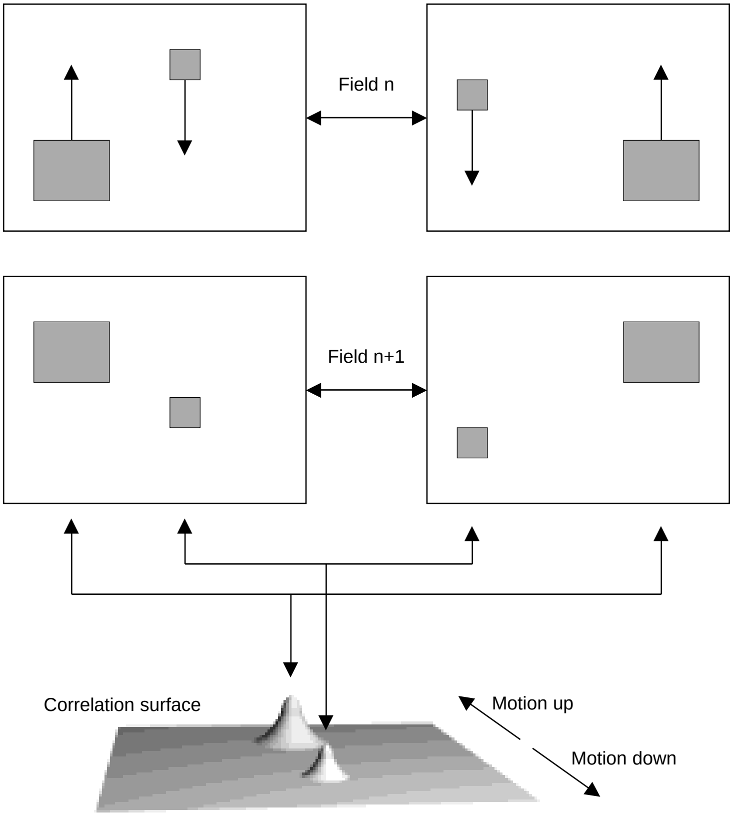


Fig 2.3.2 Phase correlation measures motion, not the location of moving objects. These two examples give the same correlation surface.

To employ the terminology of motion estimation, the phase correlation process produces candidate vectors, and a process called image correlation assigns the vectors to specific areas of the picture. In many ways the vector assignment process is more difficult than the phase correlation process as the latter is a fixed computation whereas the vector assignment has to respond to infinitely varying picture conditions.

2.4 Image correlation

Fig 2.4.1 shows the image correlation process which is used to link the candidate vectors from the phase correlator to the picture content. In this example, the correlation surface contains three peaks which define three possible motions between two successive fields. One pre-processed field is successively shifted by each of the candidate vectors and compared with the next field a pixel at a time.

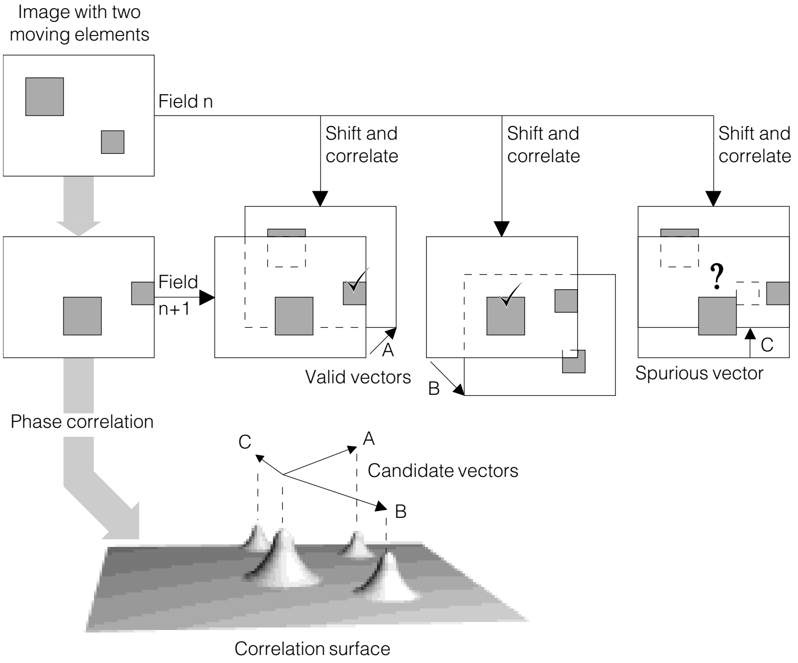


Fig 2.4.1 Image correlation uses candidate vectors to locate picture areas with the corresponding motion. If no image correlation is found, the vector was spurious and is discounted.

Similarities or correlations between pixel values indicate that an area with the measured motion has been found. This happens for two of the candidate vectors, and these vectors are then assigned to those areas. However, shifting by the third vector does not result in a meaningful correlation. This is taken to mean that it was a spurious vector; one which was produced in error because of difficult program material. The ability to eliminate spurious vectors and establish confidence levels in those which remain is essential to artifact-free conversion.

Image correlation is a form of matching because it is looking for similar luminance values in successive fields. However, image correlation is performed after the motion estimation when the motion is known, whereas block matching is used to estimate the motion. Thus the number of correlations which a block matcher must perform is very high compared to those needed by an image correlator. The probability of error is therefore much smaller. The image correlator is not looking for motion because this is already known. Instead it is looking for the outline of objects having that known motion.

2.5 Vector assignment

The phase correlation process produces candidate vectors in each window. The vectors from all windows can be combined to obtain an overall view of the motion in the field before attempting to describe the motion of each pixel individually.

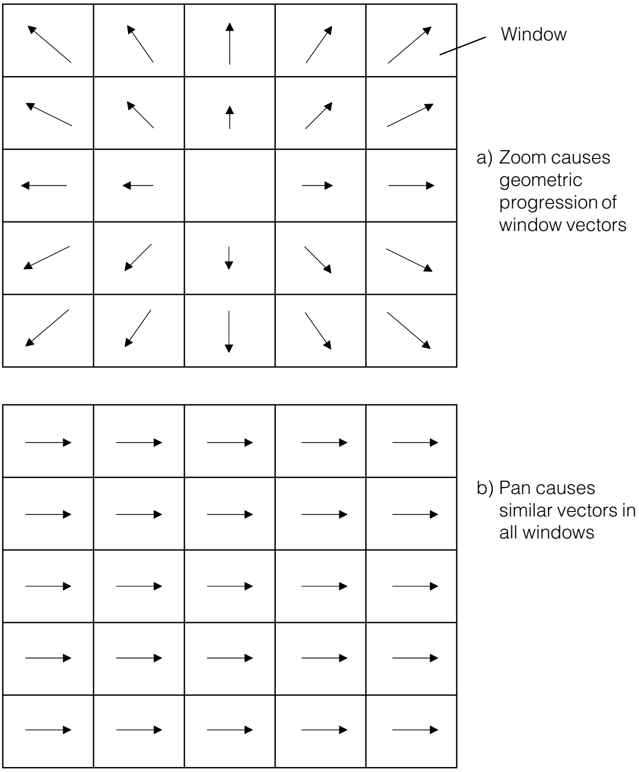

Window
a) Zoom causes geometric progression of window vectors
b) Pan causes similar vectors in all windows

Fig 2.5.1 The results of a) a zoom and b) a pan on the vectors in various windows in the field.

Fig 2.5.1a) shows that if a zoom is in progress, the vectors in the various windows will form a geometric progression becoming longer in proportion to the distance from the axis of the zoom. However, if there is a pan, it will be seen from Fig 2.5.1b) that there will be similar vectors in all of the windows. In practice both motions may occur simultaneously.

An estimate will be made of the speed of a field-wide zoom, or of the speed of picture areas which contain receding or advancing motions which give a zoom-like effect. If the effect of zooming is removed from each window by shifting the peaks by the local zoom magnitude, but in the opposite direction, the position of the peaks will reveal any component due to panning. This can be found by summing all of the windows to create a histogram. Panning results in a dominant peak in the histogram where all windows contain peaks in a similar place which reinforce.

Each window is then processed in turn. Where only a small part of an object overlaps into a window, it will result in a small peak in the correlation surface which might be missed. The windows are deliberately overlapped so that a given pixel may appear in four windows. Thus a moving object will appear in more than one window. If the majority of an object lies within one window, a large peak will be produced from the motion in that window. The resulting vector will be added to the candidate vector list of all adjacent windows. When the vector assignment is performed, image correlations will result if a small overlap occurred, and the vector will be validated. If there was no overlap, the vector will be rejected.

The peaks in each window reflect the degree of correlation between the two fields for different offsets in two dimensions. The volume of the peak corresponds to the amount of the area of the window (i.e. the number of pixels) having that motion. Thus the largest peak should be selected first.

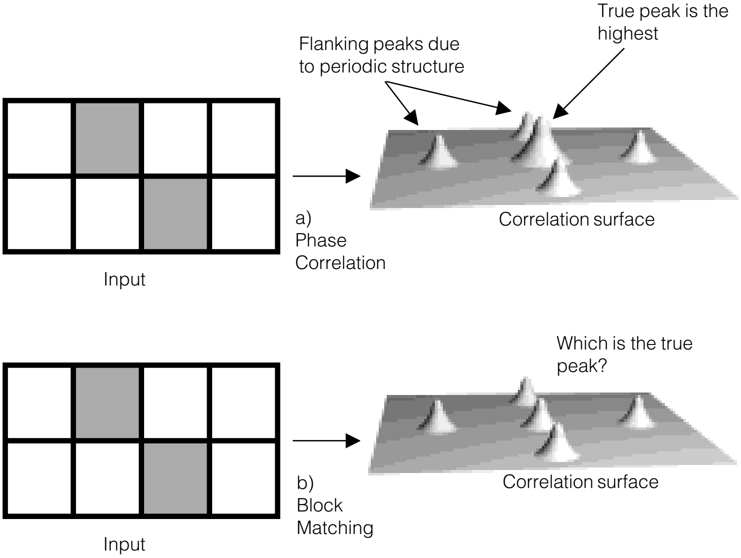


Fig 2.5.2 a Phase correlation can use the low frequency shading on the grille to identify the true motion which corresponds to the highest peak in the correlation surface

b Block matching is an amplitude domain process and is restricted to small blocks. The low frequency information is not fully employed and the correct motion cannot be established,

Periodic structures such as grilles or gratings in the input image are a traditional source of difficulty in all motion estimation systems, because they cause multiple correlations. Whilst phase correlation is not completely immune to such difficult material, it does perform better than other techniques and as a result it will fail less often on real material. There are several reasons for this. Phase correlation is computationally more efficient than other methods, and so larger blocks or windows can be used. The larger the window, the easier it is to find the correct correlation among the candidates because the phase of lower frequency components of the input video can be used. Thus phase correlation can use subtle information such as the variations in illumination of the grille itself to produce a peak in the correlation surface which is higher than the flanking peaks. Block matching cannot do this because it works in the amplitude domain and such subtle signal components result in noise which reduces the correct correlation and the false ones alike. Fig 2.5.2a) shows the correlation surface of a phase correlated motion estimator when a periodic structure is input. The central peak is larger than the flanking peaks because correlation of all image frequencies takes place there. In block matching, shown at b) the correlation peaks are periodic but the amplitudes are not a reliable indication of the correct peak. A further way in which multiple correlations may be handled is to compare the position of the peaks in each window with those estimated by the pan/zoom process. The true peak due to motion will be similar; the sub peaks due to image periodicity will not be and can be rejected.

Correlations with candidate vectors are then performed. The image in one field is shifted in an interpolator by the amount specified by a candidate vector and the degree of correlation is measured. Note that this interpolation is to sub-pixel accuracy because phase correlation can accurately measure sub-pixel motion. High correlation results in vector assignment, low correlation results in the vector being rejected as unreliable.

If all of the peaks are evaluated in this way, then most of the time valid assignments will be made for which there is acceptable confidence from the correlator. Should it not be possible to obtain any correlation with confidence in a window, then the pan/zoom values will be inserted so that that window moves in a similar way to the overall field motion.

2.6 Obscured and revealed backgrounds

When objects move, they obscure their background at the leading edge and reveal it at the trailing edge. As a motion compensated standards convertor will be synthesising moving objects in positions which are intermediate to those in the input fields, it will be necessary to deal carefully with the background in the vicinity of moving edges.

This is handled by the image correlator. There is only one phase difference between the two fields, and one set of candidate motion vectors, but the image correlation between fields takes place in two directions.

The first field is correlated with the shifted second field, and this is called forward correlation. The direction of all of the vectors is reversed, and the second field is correlated with the first field shifted; this is backward correlation. In the case of a pan, the two processes are identical, but Fig 2.6.1 shows that in the case of a moving object, the outcomes will be different. The forward correlation fails in picture areas which are obscured by the motion, whereas the backward correlation fails in areas which are revealed. This information is used by the standards convertor which assembles target fields from source fields.

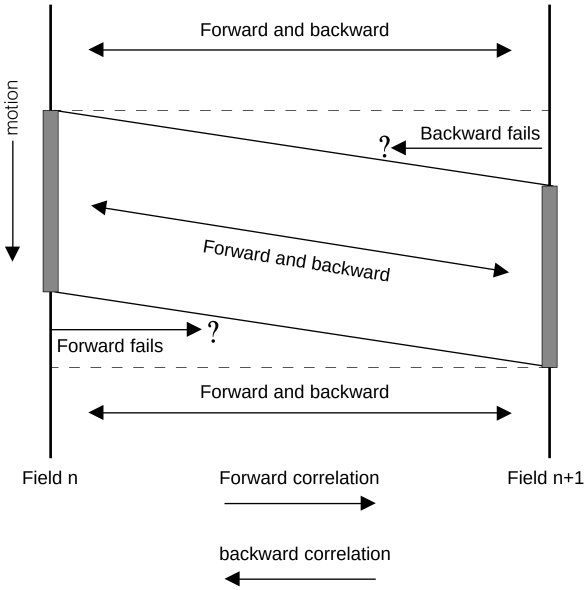


Fig 2.6.1 The difference between forward and backward correlation is due to revealed and obscured background.

Where obscuration is taking place, the convertor is told to omit pixels from earlier source fields which must be behind the new position of the moving object. The output field just ahead of the leading edge of the object is then assembled only from the information about the moving object contained in earlier source fields. Where revelation is taking place, a similar process takes place where information about the background just behind the trailing edge of the moving object comes only from later fields. In this way the background is correctly portrayed as a moving object traverses it. In order to control pixel data from earlier fields separately from pixel data in later fields two different vector fields are sent to the converter.

SECTION 3 – THE ALCHEMIST

The Alchemist is a high quality 24-point standards convertor which can operate as a stand alone unit. However, it is equipped with inputs to accept motion vectors from an optional phase correlated motion estimation unit. Operation with and without motion compensation is described here.

3.1 Standards conversion

Standards conversion is a form of sampling rate conversion in two or three dimensions. The sampling rate on the time axis is the field rate and the sampling rate in the vertical axis is the number of lines in the unblanked field. The sampling rate along the line is 720 pixels per line in CCIR-601 compatible equipment. For widescreen or high definition signals it would be higher.

Considering first the vertical axis of the picture, taking the waveform described by any column of pixels in the input standard, the goal is to express the same waveform by a different number of pixels in the output standard.

This is done by interpolation. Although the input data consist of discrete samples, these actually represent a band limited analog waveform which can only join up the discrete samples in one way. It is possible to return to that waveform by using a low-pass filter with a suitable response. An interpolator is the digital equivalent of that filter which allows values at points in between known samples to be computed. The operation of an interpolator is explained in the Appendix B.

There are a number of relationships between the positions of the known input samples and the position of the intermediate sample to be computed. The relative position controls the phase of the filter, which basically results in the impulse response being displaced with respect to the input samples. This results in the coefficients changing. As a result a field can be expressed as any desired number of lines by performing a series of interpolations down columns of pixels.

A similar process is necessary along the time axis to change the field rate. In a practical standards convertor vertical and temporal processes are performed simultaneously in a two dimensional filter.

The phase of the temporal interpolation changes as a function of the vertical position in the field owing to the different temporal slope of the fields in the input and output standards. This is shown in Fig 3.1.1.

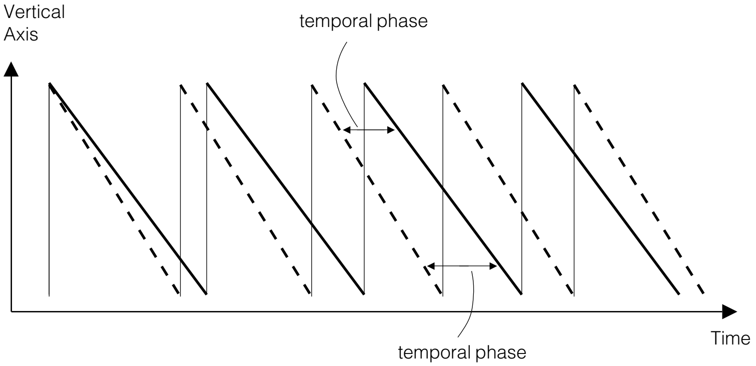


Fig 3.1.1 The different temporal distribution of input and output fields in a 50/60Hz converter.

3.2 Motion Compensation

In a motion compensated standards convertor, the inter-field interpolation axis is not aligned with the time axis in the presence of motion. Fig 3.2.1a) shows an example without motion compensation and b) shows the same example with motion compensation. In practice the interpolation axis is skewed by using the motion vectors to shift parts of the source fields.

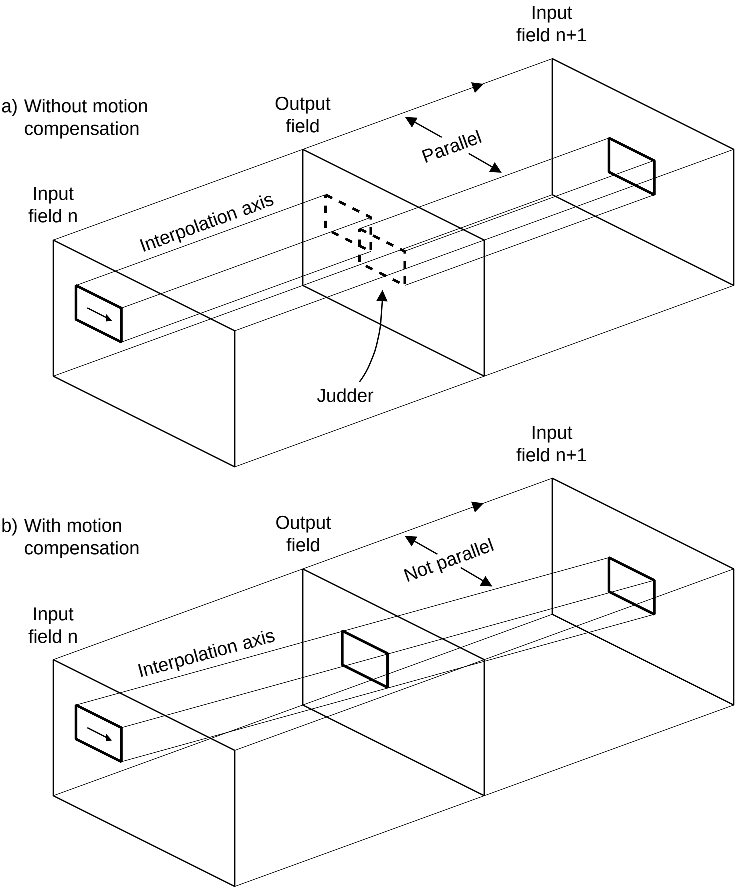


Fig 3.2.1 At a) the interpolation axis is parallel with the time axis. The axis moves as shown in b) when motion compensation is used in order to lie parallel to the axes of motion.

Shifting an area of a source field takes place in two stages. The displacement will be measured in pixels, and the value is divided into the integer part, i.e the nearest whole number of pixels, and the fractional part, i.e the sub-pixel shift. Pixels from input fields are stored in RAM which the interpolator addresses to obtain input for filtering. The integer part of the impulse response shift is simply added to the RAM address so that the pixels from the input field appear to have been shifted. The vertical shift changes the row address and the horizontal shift changes the column address. This is known as address mapping and is a technique commonly used in DVEs. Address mapping moves the image to pixel accuracy, and this stage is followed by using the sub-pixel shift to control the phase of the interpolator. Combining address mapping and interpolation in this way allows image areas to be shifted by large distances but with sub-pixel accuracy.

There are two ways of implementing the address mapping in a practical machine which are contrasted in Fig 3.2.2. In write-side shifting, the incoming source field data are sub-pixel shifted by the interpolator and written at a mapped write address in the source RAM. The RAM read addresses are generated in the same way as for a conventional convertor. In read-side shifting, the source field RAM is written as normal, but the read addresses are mapped and the interpolator is used to shift the read data to sub-pixel accuracy. Whilst the two techniques are equivalent, in fact the vectors required to control the two processes are quite different. The motion estimator computes a motion vector field in which a vector describes the distance and direction moved by every pixel from one input field to another. This is not what the standards convertor requires.

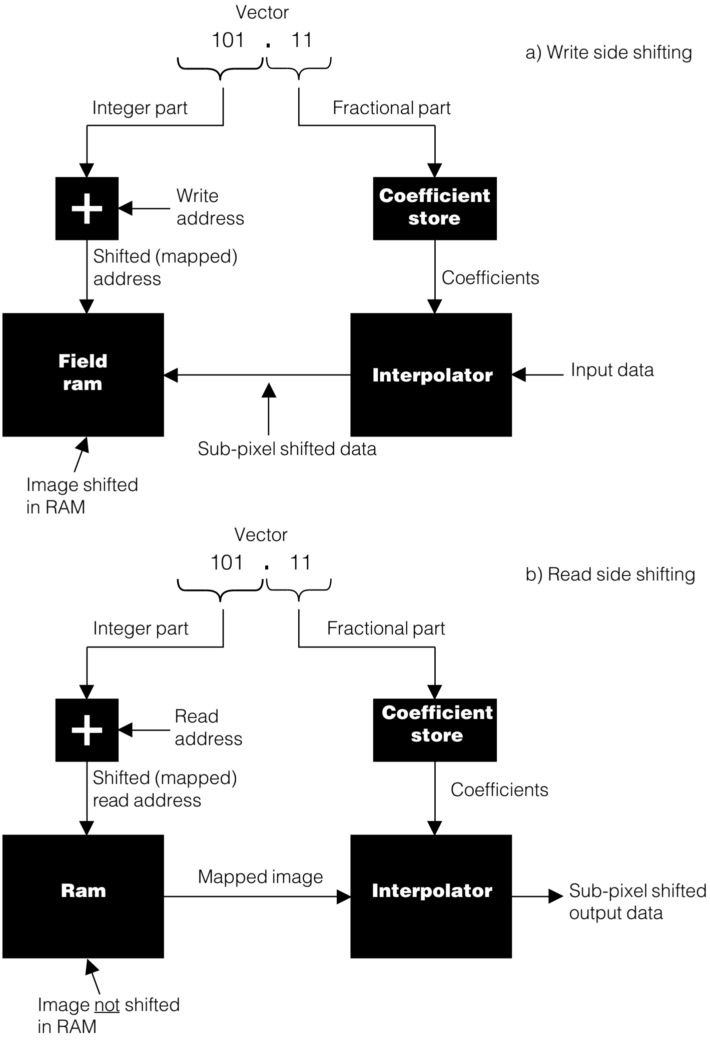


Fig 3.2.2 a Write-side modify
b Read-side modify

In write side shifting, a simple vector interpolation stage is required which computes the location of the current output field between the input fields, and uses this geometrically to proportion the motion vector into two parts. Fig 3.2.3a) shows that the first part is the motion between field A and the output field; the second is the motion between field B and the output field. Clearly the sum of these two vectors is the motion between input fields.

Whilst a conventional convertor only needs to interpolate vertically and temporally, motion compensation also requires horizontal interpolation to account for lateral movement in images. Fig 3.2.3b) shows that the motion vectors from the vector interpolator are resolved into two components, vertical and horizontal. These processed vectors are used to displace the input fields as they are being written into the field RAMs. As the motion compensation is performed on the write side of the source RAMs, the remainder of the standards convertor is nearly conventional. In read-side shifting, the motion vector processing is more complex. Fig 3.2.4 shows that it is necessary to standards convert the motion vector fields from the input field structure to the output field structure prior to geometrically proportioning the vectors to obtain the read-side shifting vectors. As the vector field is sub-sampled in real program material, the standards conversion process can introduce unwanted degradation and for this reason the Alchemist uses write-side compensation.

.

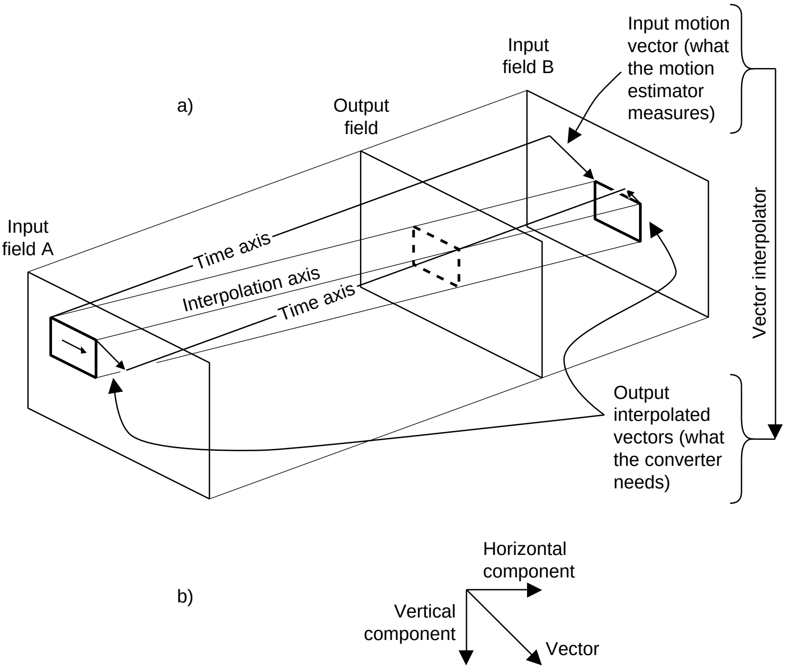


Fig 3.2.3 At a) the field to be interpolated is positioned temporally between source fields and the motion vector between them is apportioned according to the location. Motion vectors are two dimensional, and can be transmitted as vertical and horizontal components shown at b) which control the spatial shifting of input fields.

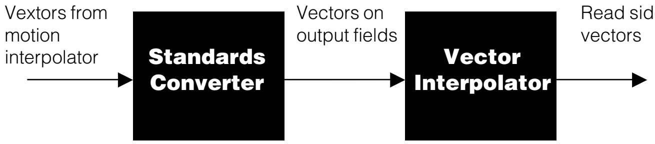


Fig 3.2.4 Read-side shifting requires the vector fields from the motion estimator to be standards converted to the output field structure before interpolation.

3.3 Handling concealment

When an object in the picture moves, it will obscure its background as was seen in section 2.6. The vector interpolator in the standards convertor handles this automatically provided the motion estimation has produced correct vectors. Fig 3.3.1 shows an example of background handling. The moving object produces a finite vector associated with each pixel, whereas the stationary background produces zero vectors except in the area O – X where the background is being obscured. Vectors converge in the area where the background is being obscured, and diverge where it is being revealed. Image correlation is poor in these areas so no valid vector is assigned.

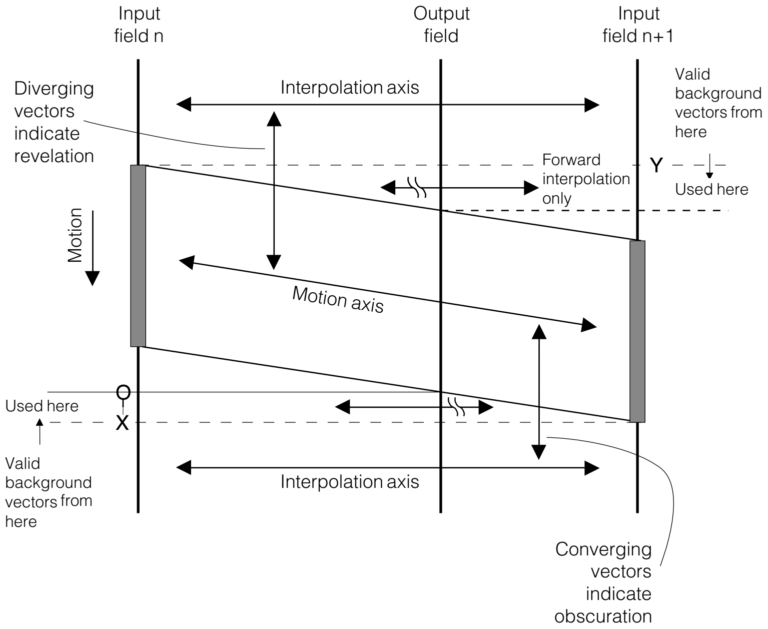


Fig 3.3.1 Background handling. When a vector for an output pixel near a moving object is not known, the vectors from adjacent background areas are assumed. Converging vectors imply obscuring is taking place which requires that interpolation can only use previous field data. Diverging vectors imply that the background is being revealed and interpolation can only use data from later fields.

An output field is located between input fields, and vectors are projected through it to locate the intermediate position of moving objects. These are interpolated along an axis which is parallel to the motion axis. This results in address mapping which locates the moving object in the input field RAMs. However, the background is not moving and so the interpolation axis is parallel to the time axis.

The pixel immediately below the leading edge of the moving object does not have a valid vector because it is in the area O – X where forward image correlation failed. The solution is for that pixel to assume the motion vector of the background below point X, but only to interpolate in a backwards direction, taking pixel data from previous fields. In a similar way, the pixel immediately behind the trailing edge takes the motion vector for the background above point Y and interpolates only in a forward direction, taking pixel data from future fields. The result is that the moving object is portrayed in the correct place on its trajectory, and the background around it is filled in only from fields which contain useful data. Clearly the convertor can only handle image data from before the output field differently to that from after the output field if it is supplied with two sets of motion vectors.

When using write side compensation RAM data are overwritten in the area of the leading edge of a moving object as this is how concealment of the background is achieved. Clearly spurious vectors will result in undesireable overwriting. The solution adopted in the Alchemist is that simple overwriting is not used, but where two writes take place to the same RAM address, a read modify write process is employed in which both values contribute to the final value. Vectors have confidence levels attached to them. If the vector confidence is high, the overwriting value is favoured. If the confidence level is low, the original value is favoured.

3.5 Comparing standards converters

Whilst many items of video equipment can be tested with conventional techniques such as frequency response and signal to noise ratio, the only way to test motion compensated standards converters is subjectively. It is, however, important to use program material which is sufficiently taxing otherwise performance differences will not be revealed.

The source of program material is important. If a tube camera is used, it will superimpose a long temporal aperture on the signal and result in motion blur. This has two effects, firstly the blur is a part of each field and so motion compensation cannot remove it. Secondly the blur may conceal shortcomings in a convertor which better material would reveal. If camera motion blur is suspected, this can be confirmed by viewing the material in freeze. The best material for testing will be obtained with CCD cameras. If shuttered cameras are available this is even better.

A good motion estimating convertor such as the Alchemist will maintain apparent resolution in the case of quite rapid motion with a shuttered CCD camera as input, whereas any inaccuracy in the vectors will reduce resolution on such material.

When viewing converted material in which the cameraman pans a moving object, there will be very little judder on the panned object even without motion estimation, so the place to look is in the background, where panning causes rapid motion. Some backgrounds conceal artifacts quite well. Grass or hedges beside race tracks do not represent a stringent test as they are featureless. On the other hand skiing and ice skating make good test subjects because in both there is likely to be sharply outlined objects in the background such as flagpoles or advertising around the rink, and these will reveal any judder. Both are likely to result in high panning speeds, testing the range of the motion vectors. Ballgames result in high speeds combined with small objects, and often contain fast pans where a camera tries to follow the ball.

Scrolling or crawling captions and credits make good test material, particularly if combined with fades as this tests the image correlator. Stationary captions with rapidly moving backgrounds test the obscuring /revealing process. The cautions outlined in section 1 regarding video from telecine machines should be heeded when selecting demonstration material. A motion compensated standards converter does not contribute judder itself, but it cannot remove judder which is already present in the source material. If use with telecine material is anticipated, it is important not to have unrealistic expectations (See section 4.4).

SECTION 4 - FURTHER APPLICATIONS

Whilst standards conversion is the topic most often associated with motion estimation, it should be appreciated that this is only one application of the technique. Motion estimation is actually an enabling technology which finds application in a wide range of devices.

4.1 Slow Motion Systems

In VTRs which are playing back at other than the correct speed, fields are repeated or omitted to keep the output field rate correct. This results in judder on moving objects, particularly in slow motion. A dramatic improvement in the slow motion picture quality has been demonstrated by the Snell and Wilcox Gazelle flow-motion technology, which is a form of motion compensated standards convertor between the VTR and the viewer in which the input field rate is variable.

Slow-motion is one of the most stringent tests of motion estimation as a large number of fields need to be interpolated between the input fields and each one will have moving objects synthesised in a different place.

4.2 Noise reduction

Noise reduction in video signals works by combining together successive frames on the time axis such that the image content of the signal is strongly reinforced whereas the random element in the signal due to noise does not. The noise reduction increases with the number of frames over which the noise is integrated, but image motion prevents simple combining of frames. If motion estimation is available, the image of a moving object in a particular frame can be integrated from the images in several frames which have been superimposed on the same part of the screen by displacements derived from the motion measurement. The result is that greater reduction of noise becomes possible.

4.3 Oversampling displays

In conventional TV displays, the marginal field rates result in flicker. If the rate at which the display is refreshed is, for example, doubled by interpolating extra fields between those in the input signal, the flicker can be eliminated. This oversampling technique is a simplified form of standards conversion which requires interpolation on the time axis. Motion compensation is necessary for the same reasons as it is in standards conversion.

4.4 Telecine Transfer

Telecine machines use a medium which has no interlace and in which the entire image is sampled at one instant. Video is displayed with scan and interlace and so there is a considerable disparity between the way the time axis is handled in the two formats. As a result judder on moving parts of the image is inevitable with a direct scanning telecine, particularly when 3:2 pulldown is used to obtain 60 Hz field rate. Proper telecine transfer actually requires a standards conversion process in the time axis which can only be done properly with motion compensation. The result is a reduction in judder. Motion estimation can also be used to compensate for film registration errors.

These applications of motion estimation are all theoretically feasible, but are not yet all economically viable. As the real cost of digital processing power continues to fall, more of these applications will move out of the laboratory to become commercial devices.

The strengths of phase correlation make it the technique of choice for all of these applications. As high definition systems become more common, needing around five times as many pixels per frame, the computational efficiency of phase correlation will make it even more important.

APPENDIX A
The Fourier transform

The Fourier transform is a processing technique which analyses signals changing with respect to time and expresses them in the form of a spectrum. Any waveform can be broken down into frequency components. Fig A1 shows that if the amplitude and phase of each frequency component is known, linearly adding the resultant components results in the original waveform. This is known as an inverse transform. In digital systems the waveform is expressed as a number of discrete samples. As a result the Fourier transform analyses the signal into an equal number of discrete frequencies. This is known as a Discrete Fourier Transform or DFT. The Fast Fourier Transform is no more than an efficient way of computing the DFT.

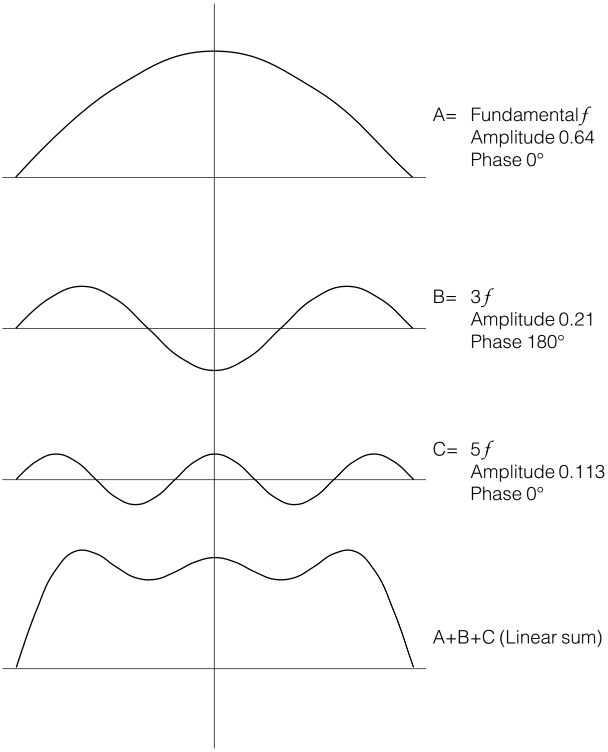


Fig A1 Fourier analysis allows the synthesis of any waveform by the addition of discrete frequencies of appropriate amplitude and phase.

It will be evident from Fig A1 that the knowledge of the phase of the frequency component is vital, as changing the phase of any component will seriously alter the reconstructed waveform. Thus the DFT must accurately analyse the phase of the signal components. There are a number of ways of expressing phase. Fig A2 shows a point which is rotating about a fixed axis at constant speed. Looked at from the side, the point oscillates up and down at constant frequency. The waveform of that motion is a sinewave, and that is what we would see if the rotating point were to translate along its axis whilst we continued to look from the side.

Fig A2a

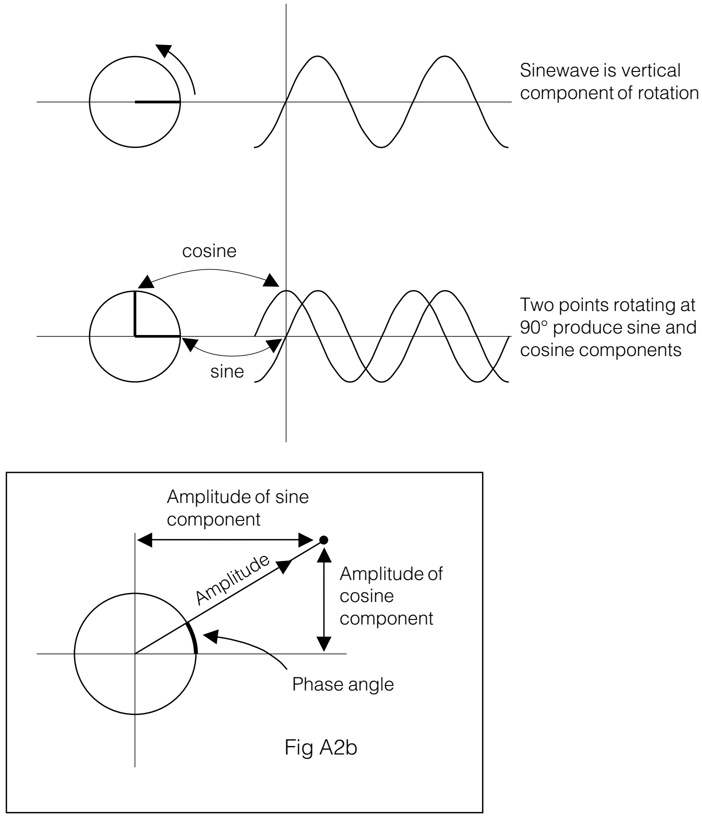

Sinewave is vertical component of rotation
cosine
sine
Two points rotating at 90° produce sine and cosine components
Amplitude of sine component
Amplitude
Amplitude of cosine component
Phase angle

Fig A2b

Fig A2 The origin of sine and cosine waves is to take a particular viewpoint of a rotation. Any phase can be synthesised by adding proportions of sine and cosine waves.

One way of defining the phase of a waveform is to specify the angle through which the point has rotated at time zero (T=0).

If a second point is made to revolve at 90 degrees to the first, it would produce a cosine wave when translated. It is possible to produce a waveform having arbitrary phase by adding together the sine and cosine wave in various proportions and polarities. For example adding the sine and cosine waves in equal proportion results in a waveform lagging the sine wave by 45 degrees.

Fig A2b shows that the proportions necessary are respectively the sine and the cosine of the phase angle. Thus the two methods of describing phase can be readily interchanged.

The Fourier transform analyses the spectrum of a block of samples by searching separately for each discrete target frequency. It does this by multiplying the input waveform by a sine wave having the target frequency and adding up or integrating the products.

Fig A3a) shows that multiplying by the target frequency gives a large integral when the input frequency is the same, whereas Fig A3b) shows that with a different input frequency (in fact all other different frequencies) the integral is zero showing that no component of the target frequency exists. Thus from a real waveform containing many frequencies all frequencies except the target frequency are excluded.

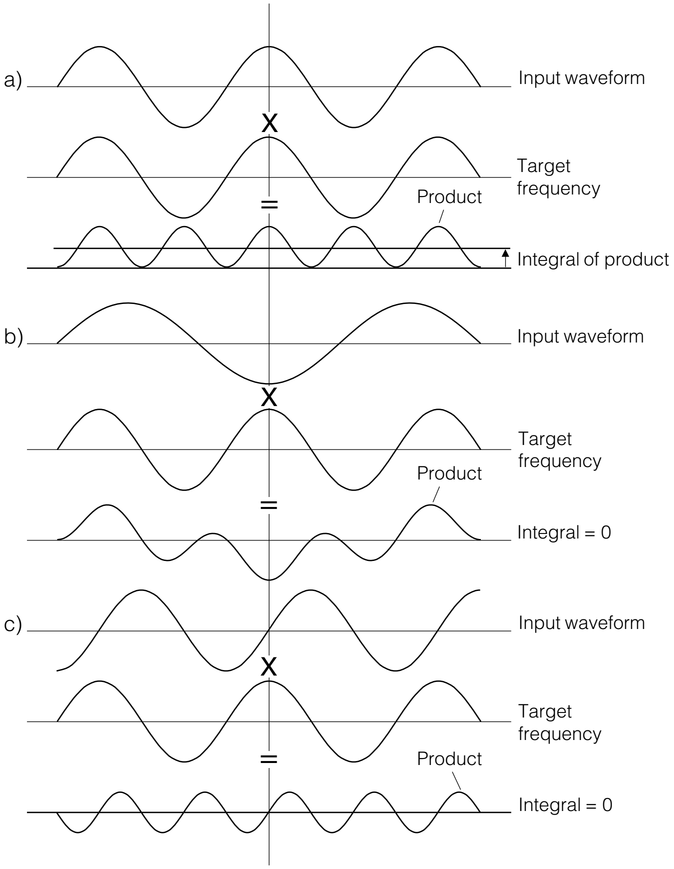


Fig A3 The input waveform is multiplied by the target frequency and the result is averaged or integrated. At a) the target frequency is present and a large integral results. With another input frequency the integral is zero as at b). The correct frequency will also result in a zero integral shown at c) if it is at 90 degrees to the phase of the search frequency. This is overcome by making two searches in quadrature.

Fig A3c) shows that the target frequency will not be detected if it is phase shifted 90 degrees as the product of quadrature waveforms is always zero. Thus the Fourier transform must make a further search for the target frequency using a cosine wave. It follows from the arguments above that the relative proportions of the sine and cosine integrals reveal the phase of the input component. Thus each discrete frequency in the spectrum must be the result of a pair of quadrature searches.

The above approach will result in a DFT, but only after considerable computation. However, a lot of the calculations are repeated many times over in different searches. The FFT aims to give the same result with less computation by logically gathering together all of the places where the same calculation is needed and making the calculation once.

The amount of computation can be reduced by performing the sine and cosine component searches together.

Another saving is obtained by noting that every 180 degrees the sine and cosine have the same magnitude but are simply inverted in sign. Instead of performing four multiplications on two samples 180 degrees apart and adding the pairs of products it is more economical to subtract the sample values and multiply twice, once by a sine value and once by a cosine value.

Fig A4 shows how the search for the lowest frequency in a block is performed. Pairs of samples are subtracted as shown, and each difference is then multiplied by the sine and the cosine of the search frequency. The process shifts one sample period, and a new sample pair are subtracted and multiplied by new sine and cosine factors. This is repeated until all of the sample pairs have been multiplied. The sine and cosine products are then added to give the value of the sine and cosine coefficients respectively. A similar process may be followed to obtain the sine and cosine coefficients of the remaining frequencies.

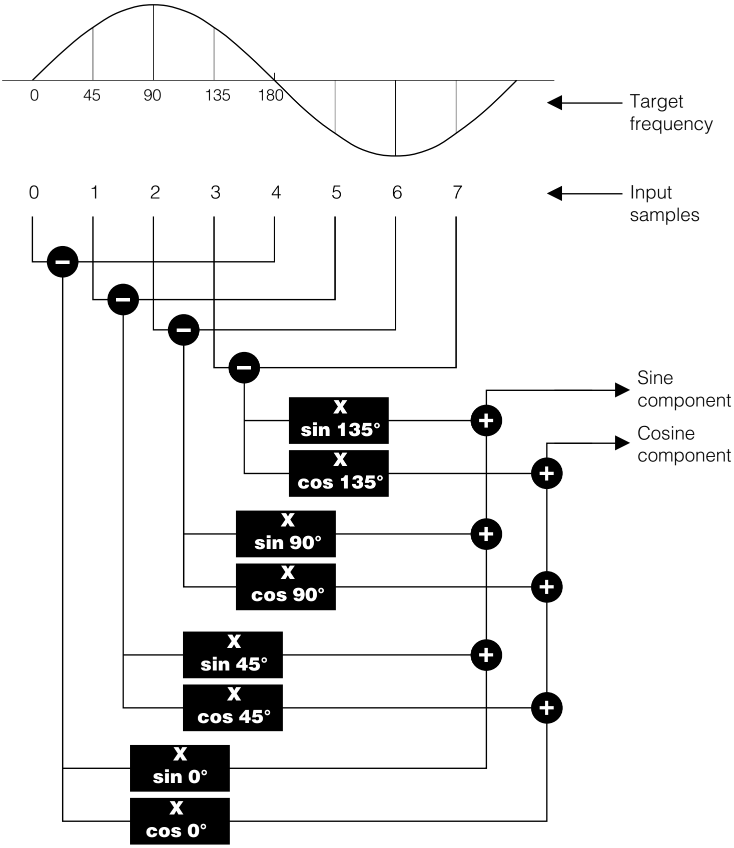


Fig A4 An example of a filtering search. Pairs of samples are subtracted and multiplied by sampled sine and cosine waves. The products are added to give the sine and cosine components of the search frequency.

In practice some tricks are used to further reduce computation, but as the result is the same it is not necessary to go into such details here. As a result of the FFT, the sine and cosine components of each frequency are available. For use with phase correlation it is necessary to convert to the alternative means of expression, i.e phase and amplitude.

The number of frequency coefficients resulting from a DFT is equal to the number of input samples. If the input consists of a larger number of samples it must cover a larger area of the screen, but its spectrum will be known more finely.

Thus a fundamental characteristic of such transforms is that the more accurately the frequency and phase of a waveform is analysed, the less is known about where such frequencies exist on the spatial axis.

APPENDIX B

Interpolation

Interpolation is the process of computing the values of samples which lie between existing samples on the original analog waveform. It is thus a form of sampling rate conversion. One way of changing the sampling rate is to return to the analog domain in a DAC and then to sample at the new rate. In practice this is not necessary because the process can be simulated in the digital domain. When returning to the analog domain a suitable low pass filter must be used which cuts off at a frequency of one half the sampling rate.

Fig B1 shows that the impulse response of an ideal low-pass filter (LPF) is a sinx/x curve which passes through zero at the site of all other samples except the centre one. Thus the reconstructed waveform passes through the top of every sample. In between samples the waveform is the sum of many impulses. In an interpolator a digital filter can replace the analog filter.

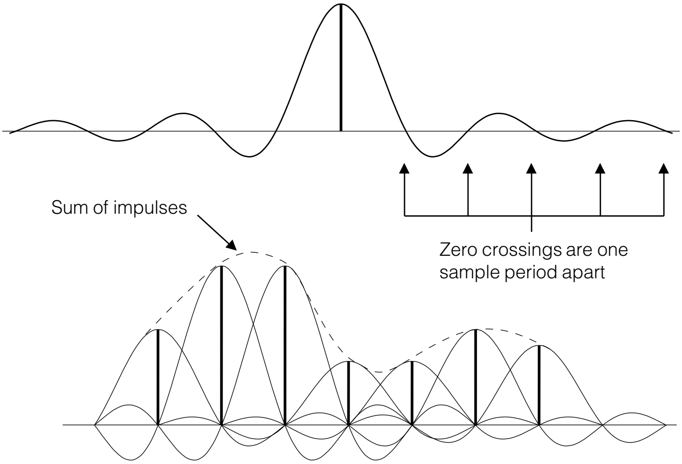


Fig B1 Low pass filtering in a DAC results in the impulses from individual samples adding to produce an analog waveform

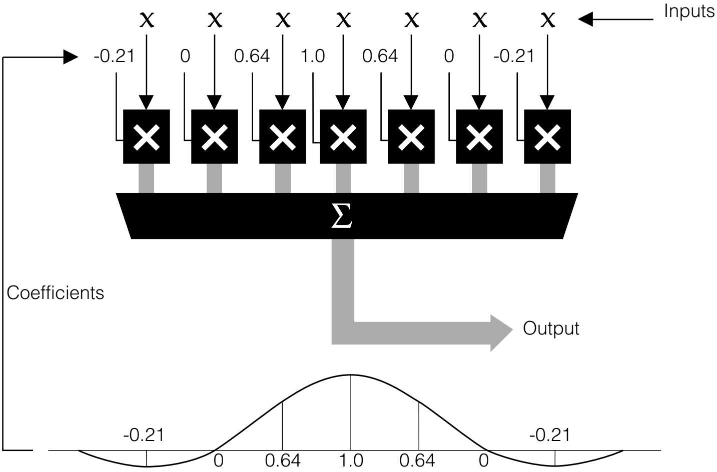


Fig B2 A digital filter can be made with a linear phase low pass impulse response in this way. As a unit input sample shifts across the filter, it is multiplied by various coefficients which produce the impulse response.

Fig B2 shows a phase linear digital filter which has a low-pass impulse response. As an input sample shifts across the filter, it is multiplied by different coefficients to create the impulse response.

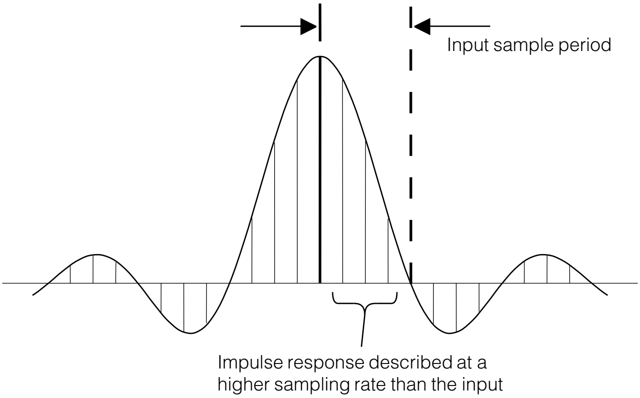

Input sample period
Impulse response described at a
higher sampling rate than the input

Fig B3 An impulse response can be sampled more often so that coefficients between input samples are known

Fig B3 shows that the impulse response can be sampled more often so that the output is available at a higher sampling rate.

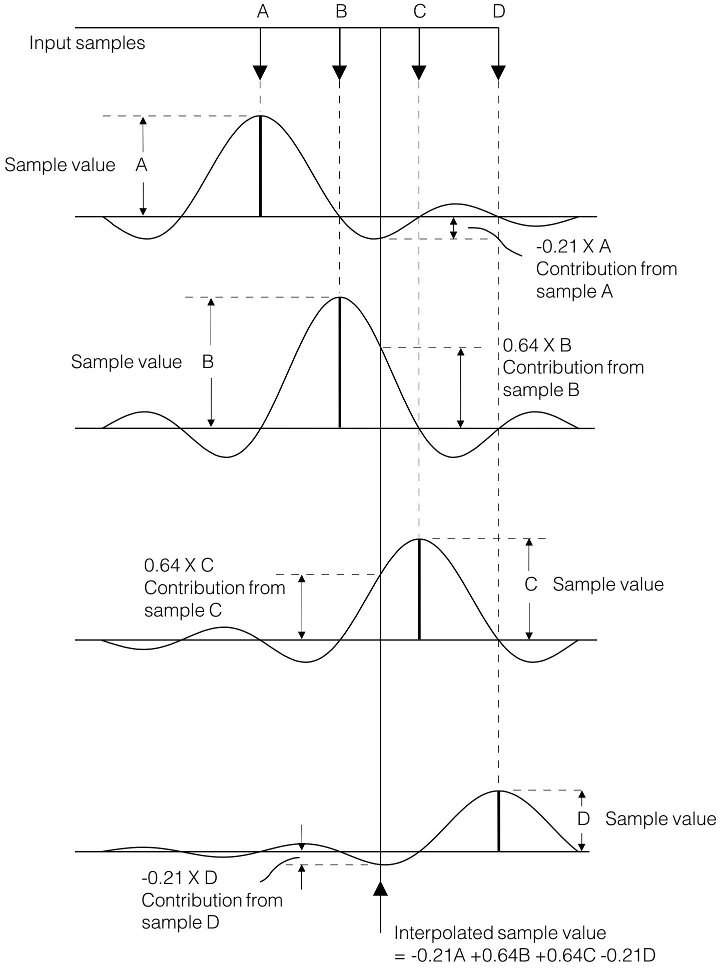


Fig B4 Interpolation. The contribution of each input sample at the corresponding distance from the required output sample is computed. All of the contributions are added to obtain the interpolated value.

Fig B4 shows the process needed to interpolate to an arbitrary position between samples. The location of the output sample is established relative to the input samples (this is known as the phase of the interpolation), and the value of the impulse response of all nearby samples at that location is added. In practice the coefficients can be found by shifting the impulse response by the interpolation phase and sampling it at new locations. The impulse will be sampled in various phases and the coefficients will be held in ROM. A different phase can then be obtained by selecting a different ROM page. Alternatively the coefficients may be calculated dynamically by a suitable processor.

Published by Snell & Wilcox Ltd.
Durford Mill
Petersfield
Hampshire
GU13 5AZ
Tel: +44 (0) 703 821188
Fax: +44 (0) 703 821199

SNELL & WILCOX