Improved road surveillance through digital enhancement

Gabriel Scarmana
BSurv (Hons), MSurv&MapSc, PhD

Abstract

Video imagery plays an important role in many Intelligent Transportation Systems (ITS) applications, such as traffic monitoring, traffic law reinforcement, driver assistance, and automatic vehicle guidance. These systems installed in either outdoor environments or vehicles often suffer from image quality and/or instability. These quality and instability factors can be rectified by way of video stabilisation and multi-frame image enhancement techniques.

In this context, the aim of video stabilisation is to create a new video sequence where the motions (i.e. rotations and translations) and scale differences between frames (or parts of a frame) have effectively been removed. The stabilisation effect is obtained using the information extracted from the video sequence itself, with no need for additional hardware or knowledge about camera physical motion.

Since a video sequence usually contains a large overlap between successive frames, regions of the same scene are sampled at different positions. This sampling can be combined via multi-frame image enhancement processes to achieve images of improved quality.

Introduction

When a scene is imaged with a hand-held, a vehicle-mounted video camera or a surveillance video camera, the result is a distorted representation of the view. However, under certain conditions, it is possible to extract and merge multiple stabilised video frames to produce an enhanced composite of a region of interest. This paper describes a method which exploits existing video stabilisation processes followed by a multi-frame image enhancement technique.

Video stabilisation technology is used to avoid the loss of visual quality by reducing unwanted shakes and jitters without influencing moving objects or intentional camera panning. Unstable video imagery is typically caused by unwanted hand jiggling, instabilities associated with adverse windy conditions, vibration caused by passing objects (e.g. trucks and aircraft) and earthquakes. In all these cases, a video stabilisation process ensures superior visual accuracy and quality.

The frames extracted from a stabilised video sequence can be combined or fused to obtain higher quality images. This can be achieved by way of a popular multi-frame image processing technique referred to as image stacking or shift-and-add method.

This technique is commonly used for obtaining high quality images from a number of short exposure images with varying image shifts. It has been used in astrophotography for several decades, and is the basis for the image-stabilisation feature on some cameras. The method involves calculation of the differential shifts between images which are then shifted back to a common centre and added together. The result is an image with higher resolution than any of the original input images.

In summary, the reconstruction of a sharper and/or improved imagery from an unstable video sequence is achieved in four steps:

1. Stabilising the video sequence
2. Extracting from the stabilised video the frames outlining the scene of interest
3. Combining or stacking the stabilised frames
4. Recovering the desired enhanced image composite.

As shown in Figure 1 (overleaf), traffic cameras are often mounted at the top of long slender steel poles so as to maximise the field of view. During adverse windy conditions movements of these cameras lead to many unwanted motion effects, thus reducing video quality. The scope of this article has been restricted to the manipulation of the electronic data once it has been captured by the video camera. It may prove beneficial if further research occurred into the dynamics of the mounting pole to ensure that the pole design was optimised for reduced vibration.
Video stabilisation

To achieve video stabilisation there are two major approaches - hardware and digital techniques. The hardware approach may comprise of optical, electronic, and mechanical systems whereas digital techniques use video image processing methods.

Optical stabilisation - compensation for vibrations is managed by varying the optical path to the sensor using a floating lens element moved orthogonally to the optical axis of the lens. Vibrations are revealed by sensors and then mechanically controlled lenses instantly compensate the jitter with a correction movement before visual data is recorded. Therefore, the system response is synchronised with the vibration. Since no manipulation is done on visual data, optical stabilisation preserves image quality. Unfortunately, the high cost of optical stabilisation prevents its inclusion in low-end digital/video cameras.

Electronic stabilisation - electronic systems are used to control the stabilisation process. If the camera sensors detect a camera shake, it responds by slightly moving the digital image so that the image virtually remains in the same position on the image sensor. This movement is obtained by readdressing the area of the image sensor which is read by the capturing chip. Since the used area is small, image motion induces blur and graininess with consequent image degradation. This issue can be solved using oversized sensors or by digitally zooming the image, however, both these approaches produce some loss of resolution.

Mechanical stabilisation — camera motion is resisted by built-in gyroscopes. The gyroscope wheels, occupying opposed axes to each other, spin with high speed and physically resist camera vibrations, acting like an invisible tripod. Once the camera motion is detected, the sensor is countermoved to avoid vibrations and to obtain steady images and jitter-free panning effects.

Digital stabilisation — unlike hardware stabilisation solutions, digital video stabilisation is typically considered to contain three successive steps:
1. Motion estimation
2. Motion filtering
3. Image composition.

Motion estimation is attained by way of stepping through a video event one frame at a time and estimating the motion parameters, namely, two shifts, one rotation angle and a zoom factor. With the motion estimated between all pairs of neighbouring frames, each frame is warped to align with the last frame of the sequence. Figure 2 and 3 above illustrate this principle.
Figure 2 shows a sequence of frames from an un-stabilised video sequence while Figure 3 shows the same frames after the stabilisation process. Note that in the stabilised sequence cropping has occurred and part of the scene is lost. Only those parts of the scene which are common to all frames in Figure 2 are used in the final representation in Figure 3.

The knowledge on translations, rotations scale and/or zoom differences between the video frames is used in the motion filtering step to determine the absolute motion parameters which track camera movements frame by frame. An issue in camera motion filtering is reducing image blurriness, which is also called motion deblur. Motion blur is caused by a moving scene point that spreads out several pixel locations during the exposure period of the camera sensor.

Finally, the image composition step corrects the frames in order for a stabilised sequence. During this phase, there is the problem of some images being cropped as demonstrated in Figure 3. Filling up artificially those missing image areas is called image or video inpainting. For a complete development of the formulations and theoretical background related to the computation of these motion parameters the reader is referred to the references (1, 2).

Of course, trying to motion stabilise video sequences manually would take hours just to process a few seconds of video. For this purpose, existing video stabilising software referred to as Deshaker\textsuperscript{1} was considered and tested in this work. Deshaker works by using motion estimation techniques to determine what has moved since the previous frame. Since the movement can be caused by the camera moving or by the object of interest moving, the stabilisation of the video sequence requires the user to set the Deshaker controls so as to determine whether the process is to stabilise movement caused by the camera or by the object of interest within the captured scene.

Deshaker runs under the Windows operating system and is available as a free download. It runs as a filter within video processing software called VirtualDub\textsuperscript{2}. This software is also available as a free download. VirtualDub is currently only available in a 32-bit version and can read a number of video file formats including AVI or MPEG-1. The software corrects for panning, rotation and zoom, each adjustable separately and has an automatic inpainting option. Deshaker also allows changing the image resolution during the stabilisation process thus avoiding the need to run the video through another application.

Image stacking

When attempting to capture a traffic scene from a surveillance camera, the atmosphere distorts the scene due to smoke (dust), temperature/density gradients and moisture gradients in the form of mist or fog. A good example of this is the heat haze seen on a very hot day. These effectively change the refractive index of the air causing light to bend resulting in images being blurry effectively creating a substantial amount of ‘noise’.

\textsuperscript{1}http://www.guthspot.se/video/deshaker.htm
\textsuperscript{2}http://downloadvirtualdub.com/
Hence even the best quality image possible will still be blurry. However, by extracting several hundred images, ‘registering’ (aligning their centres) then ‘stacking’ them (i.e. adding them all together pixel-by-pixel then dividing each pixel by the number of images) the blurriness diminishes thus producing a higher quality image or composite which will contain more details than any of the individual original images.

In order to capture and generate smooth, detailed and noiseless images, many identical or semi-identical images are required to be combined or ‘stacked’. The theory is that the more images taken and stacked, the smoother and cleaner the image becomes. In addition, this process can:

1. Reduce artefacts created by compression methods
2. Reduce image noise without compromising details in the image
3. Effectively freeze atmospheric distortions while retaining image integrity
4. Increase the dynamic range of an image.

The dynamic range represents the difference between the brightest possible recordable pixel values and the dimmest possible recorded pixel values. Values greater than the brightest possible value saturate (and are therefore ceilinged as the brightest possible recordable value instead of their actual value). On the other hand, values dimmer than the dimmest possible pixel value simply drop off the bottom and are recorded as 0.

Image stacking works on the assumption that the noise in an image is truly random. This way, random fluctuations above and below actual image data will gradually even out as more and more images are stacked. In order to perform a stacking operation the images should be first accurately aligned.

In this work, the process of stacking images was carried out using a freeware program called Registax\(^1\). This program is a correction tool for images which allows improvements in contrast, brightness, colour and quality of pictures. The alignment of images requires that the user selects features in the image as a reference for the grading and alignment of the other images of the same scene. This is done by placing square regions of various sizes manually with the mouse on a single frame (that is, the reference frame).

**How many frames?**

Is it better to use a smaller number of higher quality frames or a larger number of lower quality frames? For instance, out of 1000 frames, should the best 30% of images for a total of 300 images be used, or would it be better to take 600 frames which are the best 60%? Tests indicate that more frames are better. 600 frames out of 1000 is noticeably better (less noisy and more detailed) than just 300 frames out of the same 1000 originals. In other words, a greater number of frames is more important than a higher quality cut-off. At 32 frames per second, one minute of video can generate approximately 2000 frames. However, frame rates above 15 frames per second cause most video cameras to heavily compress the image data thus compromising even further the quality of the imagery.

This means the required number of frames for enhancing a region of interest from a video sequence will generally depends on the amount of compression applied, the amount of noise and the accuracy of the image registration process. The reconstruction of a higher resolution image with the minimum number of low-resolution video frames may always improve the scene visually, but it should not be expected to always achieve a high accuracy.

\(^1\) http://www.astronomie.be/registax/download.html

![Figure 4 - Sequential scene of a 60km/h vehicle captured with an unstable hand held video camera.](image-url)
Test case

In this experimental work, a video sequence is filmed using an off-the-shelf video camera. The camera captured a road scene with the purpose of identifying a vehicle and a person within the scene. Sudden random movements were deliberately induced to the video camera so as to emulate the real life movement of traffic surveillance cameras located along motorway corridors.

A set of 500 frames were extracted from 60 seconds of colour video footage of a road scene taken from an overpass bridge using an off-the-shelf video camera. The video resolution was set to a format of 1280x720. Two tasks were considered in this test case:

1. The first test related to the moving car where the objective was to enhance the vehicle’s registration plate. The car approached the camera at a speed of approximately 60km/h from a distance of 350m (Figure 4, 5).

2. The second task related to enhancing the image of a person’s face standing near a bus shelter from a distance of 130m (Figure 6).

Figure 5 – Left image is original. Right image is clearer after combining 300 frames. Notice details of the car plate.

Figure 6 – Bottom left image is original. Top left is clearer image after combining 300 frames.
Figure 5 illustrates the two images of the vehicle before and after stabilisation and enhancement. Since the proposed image enhancement technique works with images where the objects of interest are approximately of the same scale and dimensions, a rescaling/warping technique was required for the case of the moving vehicle. In this way all the extracted images could be readily compared, registered and finally combined. The reconstructed image presents more details than any of the original frames extracted from the original video sequence.

Using the same settings as in the first test, similar results were obtained for the person of interest in Figure 6. The bottom left shows an original frame of the person extracted from the footage and the enhanced composite shown on the top left.

In this experiment all the video imagery was captured in colour. Colour images can be considered as three separate images containing red, green and blue components (RGB). Each of these components or channels were enhanced independently and then fused to produce an enhanced colour image with enhanced resolution. The underlying premise is that for any colour image sequence, the motion between adjacent frames for each colour channel should be exactly the same.

Conclusion

The problem of image resolution enhancement from unstable video can be addressed by exploiting multiple frames of interest that offer unique perspectives of a specific scene of interest. The focus here was to exploit frame-to-frame motions that may result from line-of-sight jitter of a sensor (i.e. a video camera). Exploiting these motions requires accurate estimates of them. In this context, a method for extracting higher quality imagery from a compressed, noisy and distorted video sequence using a multi-frame image enhancement approach has been presented. The process uses two techniques in sequence - video stabilisation and image stacking. This process does not rely on control points for the accurate registration and combination of the images.

The registration or matching methodology and subsequent use of the proposed enhancement technique may lead to a general approach to the problem of generating a higher resolution image from compressed, distorted and noisy video images of the same scene. The application and effectiveness of the enhancement process has been demonstrated in a test and refinements to the technique are being undertaken to increase the accuracy of the results.

This may extend the range of applications which could benefit from utilising this device independent image enhancement process, possibly adapting this method to a generalized scheme whereby both sensors and objects of interest are dynamic and the illumination is non-uniform.

It is the author’s belief that this enhancement technique may be ideal for surveillance systems based on video cameras located along busy motorways.

References