16th Australasian Remote Sensing and Photogrammetry Conference
27-29 August, 2012
Melbourne, Australia

Image enhancement from a stabilised video sequence

Gabriel Scarmana
University of Southern Queensland
Australia

Key words: Multi-frame image enhancement, video stabilisation

Abstract

The aim of video stabilisation is to create a new video sequence where the motions (i.e. rotations, translations) and scale differences between frames (or parts of a frame) have effectively been removed. These stabilisation effects can be obtained via digital video processing techniques which use the information extracted from the video sequence itself, with no need for additional hardware or knowledge about camera physical motion.

A video sequence usually contains a large overlap between successive frames, and regions of the same scene are sampled at different positions. In this paper, this multiple sampling is combined to achieve images with a higher spatial resolution. Higher resolution imagery play an important role in assisting in the identification of people, vehicles, structures or objects of interest captured by surveillance cameras or by video cameras used in face recognition, traffic monitoring, traffic law reinforcement, driver assistance and automatic vehicle guidance systems.

1 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 so as to produce an enhanced composite of a specific region of interest (ROI). To this end, 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 of a video footage taken with an image/video capturing device without influencing moving objects or intentional

camera panning (Gonzalez and Woods, 2007). Unstable images' are typically caused by undesired hand jiggling, instabilities associated with static sensors under, for example, adverse windy conditions, vibration caused by passing objects (e.g., trucks and airplanes) and earthquakes. In all these cases a video stabilisation process ensures superior visual quality and stable video footages. The video stabilisation principle is graphically illustrated in Figure 1.

(a)

(b)

Figure 1 – Frames extracted from an un-stabilised video sequence (a) and the same frames as extracted after stabilisation (b). Note that in the stabilised sequence cropping has occurred and part of the scene is lost.

The frames extracted from a stabilised video sequence can often be combined to obtain higher quality imagery of improved spatial resolution. In this work, a popular multi-frame image enhancement technique referred to super-resolution (SR) was chosen for this task. SR is a technique which uses multiple low-resolution frames of the same scene to achieve a higher resolution image of that scene. It works only if the frames are shifted by fractions of a pixel from each other. Details on how to super-resolve a set of low-resolution images of the same scene is described in the ensuing sections.

In summary, recovering a sharper and/or improved image from an unstable video sequence is achieved in four steps:

(1) Stabilisation the video sequence using digital techniques
(2) Extraction the frames depicting the same scene of interest
(3) Combination or fusion the stabilised frames to fractional pixel accuracy
(4) Recovering and displaying the enhanced image composite

2 Video stabilization

To achieve video stabilisation there exist two major approaches: (a) hardware techniques and (b) digital techniques. A hardware approach comprises optical, electronic, and mechanical systems whereas digital techniques use video/image processing methods.

In optical stabilisation (Wang, 2003), vibrations are compensated 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. Thus, the system response is synchronized with the vibration. Since no manipulation is done on visual data, optical stabilisation preserves image quality. Unfortunately, high cost of optical stabilisation devices prevents from including them in low-end digital cameras.

Electronic stabilisation uses an electronic system to control the stabilisation process (Bovik, 2007). If the system detects through its sensors a camera shake, it responds by slightly moving the image so that it virtually remains in the same position on the image sensor. This movement is obtained by re-addressing the area of the image sensor which is read by the capturing chip (ref.). 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 (Tomasi and Kanade, 1991).

In mechanical stabilisation, camera motion is detected by gyroscopes. The gyroscopic 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 counter-moved to avoid vibrations and to obtain clear, steady images and jitterless panning effects (Batur and Flinchbaugh, 2006).

Unlike hardware stabilisation solutions, digital video stabilisation is typically considered to contain three successive steps: (a) motion estimation, (b) motion filtering, and (c) image composition. Motion estimation is attained by way of stepping through a video event one frame at a time and estimating the motion parameters (Marcenaro et. a.l, 2001). The modelling of the motion between two sequential frames can be estimated with a two-dimensional linear model which usually provides a trade-off between effectiveness and complexity (Liang et al., 2004). This model describes inter-frame motion using four different parameters;

namely, two shifts, one rotation angle and a zoom factor. The model associates a point (x_i, y_i) in frame I_n with a point (x_f, y_f) in frame I_{n+1} with the following transformation:

$$\begin{cases} x_f = x_i \lambda \cos\vartheta - y_i \lambda \sin\vartheta + T_x \\ y_f = x_i \lambda \sin\vartheta + y_i \lambda \cos\vartheta + T_y \end{cases}$$

where λ is the zoom parameter, θ the rotation angle, T_x and T_y respectively *X*-axis and *Y*-axis shifts. In order to estimate four transformation parameters, four different linear equations are required. Thus, only two couples of features allow for the system to find a solution. Since features can be affected by noise, it is useful to apply a linear least squares method on a set of redundant equations (Auberger and Miro, 2005). For a comprehensive development of the formulations and theoretical background of affine transformations models as applied to video processing and stabilisation the reader is referred to Bovik (2009).

The knowledge of 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 (Bovik, 2007). Motion blur is caused by a moving scene point that spreads out several pixel locations during the exposure period of the sensor.

Finally, the image composition step corrects the frames in order to obtain the stabilised sequence (Morimoto and Chellappa, 1996). During this phase the images often suffer from the problem of some areas being cropped and/or trimmed. Filling up those missing image areas is called image or video inpainting (Matsushita et al., 2006).

Inpainting can be accomplished in several ways. The simplest solution is to fill this data with a predefined colour. By applying a scaling function to the image, these borders can be kept to a minimum. A second method is to keep the old information of the previous video frame(s). The third one is to warp the old information (Tico et. al. 2006). The last method is preferred over the two others since this one gives a more natural viewing result for the same computational effort as the second and a much better result than the first. Optionally, by displaying the stabilised video in a larger frame and preserving the information of previous frames for a longer time, one can visualize a history trail (Battiato, 2007).

Inpainting works well for static and planar scenes, but produces visible artifacts for dynamic or non-planar scenes. If inpainting is impractical trimming of the video footage may be considered thus displaying only the ROI or the portion of video that appears in all or in the majority of the frames. Moreover, sometimes

due to severe camera-shake, there might be no common area among neighboring frames. In this instance, a reduction of the frame rate may be considered to accommodate only those frames that contribute to the scene of interest.

Trying to motion stabilise a video footage manually would take hours just to process a few seconds of video. For this purpose a video stabilising software referred to as *Deshaker* (http://www.guthspot.se/video/deshaker.htm) was considered and tested in this work. *Deshaker* works by using motion estimation algorithms similar to what are used for MPEG2 video encoding (Bovik, 2009) to determine what has moved since the previous frame.

With *Deshaker*, the stabilisation process tracks salient features common to all images of the footage and uses these as anchor points to cancel out all perturbations and/or motions relative to the mages. This procedure, however, must be bootstrapped with knowledge of where such a salient feature lies in the first video frame. *Deshaker* works without any such a priori knowledge. It instead automatically searches for the "background plane" in a video sequence, and uses its observed distortion to correct for camera motion.

However, 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 in VirtualDub, a video processing software which can also be downloaded for free. At present, the software is currently available only in a 32-bit version and can read a number of video formats including AVI or MPEG-1 files. The software corrects for panning, rotation and zoom, each adjustable separately and has an automatic border-fill option, and allows changing the image resolution during the stabilisation process.

Once all the extracted video frames are stabilised, processed and warped to a common orientation using the methodology described in the previous paragraphs, the next step is to register and map all the extracted images to a regular reference frame, using the fractional shifts existing (if any) among them. Accurate fractional shifts are necessary for the correct combination and/or mapping of these mages so as to obtain the aspired higher resolution image via super-resolution.

3 Multi-frame super-resolution

The term super-resolution (SR) refers to the process of obtaining higher-resolution images from several lower-resolution ones. The quality improvement is

caused by fractional-pixel displacements between images. That is, the ROI is sampled at more locations than originally detected by the sensor array (Farsiu et al. 2004).

Hence, super-resolution allows overcoming the limitations of the imaging system (resolving limit of the sensors) without the need for additional hardware. In this work, the multiple under-sampled and degraded images of the same scene are extracted from stabilised image sequences. The additional information available in these frames makes it possible the reconstruction of visually superior frames with higher resolution (i.e. more pixels).

SR image reconstruction may have the following effects on the final image composite: (1) reduce artefacts created by compression (2) reduce image noise without compromising details in the image (3) effectively freeze atmospheric distortions while retaining image integrity and (4) increase the dynamic range of an image (Zhouchen, and Heung, 2004). The dynamic range represents the difference between the brightest possible recordable pixel values and the dimmest possible recorded pixel values.

The majority of the literature on super-resolution in the spatial domain describes the use of three basic steps:

- Estimation of the shifts among the different low-resolution images at a fractional pixel level (sometimes referred to as image-to-image registration or image matching),
- Projection of the pixels of the low-resolution images onto a higher resolution grid using the fractional pixel values detected, and,
- Interpolating or solving sets of equations derived from the geometric relationships existing between the low-resolution pixels and the high-resolution pixels.

The fractional pixel registration between two images of the same scene is derived from image matching (Wold and DeWitt, 2000). The image registration technique used in this work matches the intensity values of two digital images, while simultaneously detecting, and locating, any small geometric differences that exist between the two images.

The registration is based on a least squares area based matching technique which can overcome difficulties arising from radiometric differences in the images being matched to achieve fractional pixel accuracies of approximately 0.1 pixels. The reader is referred Pilgrim (1991) for the theory and background behind this process.

The matching process allows images to be registered without using control points in the registration procedure. For a correct detection of the shifts or offsets between two images, the images must contain some features that make it

possible to match two low-resolution images. Very sharp edges and small details are most affected by aliasing, so they are not reliable to be used to estimate these shifts. Uniform areas are ineffective, since they are translation invariant. The best features are slow transitions between two areas of grey values.

These areas are generally unaffected by aliasing and such portions of an image need not to be detected specifically, although their presence is very important for an accurate registration result. Hence, prior to matching and/or regsitering two or more images of the same scene it is recommended to remove details affected by aliasing by applying equally a low-pass filter to the images. The purpose of a low-pass filter is to “smooth” sharp edges, small details, sudden changes of intensity values and distortions created by compression processes (Vandewalle et al., 2005).

4 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% 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.

The more images available to begin with, the higher the quality cut-off that can be used while still having a relatively large number of frames to fuse. This means extracting (if available) at least 1000 original frames if the best 40% are required. On the other hand, if the best 10% is the requirement then 4000 original frames should be extracted. At 30 frames per second (fps), 4000 frames requires over two minutes of footage.

5 Reconstructing a higher resolution image

Once all the low-resolution images have been stabilised and registered to a fractional-pixel level, they are projected or mapped on a uniformly spaced high-resolution grid (see Figure 2). In the idealized super-resolution set-up of Figure 2 the images (b)-(d) are taken with fractional-pixel shifts of half a pixel in the horizontal, vertical and diagonal directions in relation to image (a).

Their pixels can then be interleaved to generate an image with a magnification factor equal to 4, that is, the image contains 4 times more real pixels than any of the low-resolution images.The horizontal, vertical and diagonal directions in relation to image (a). Their pixels can then be interleaved to generate an image with a magnification factor equal to 4, that is, the image contains 4 times more real pixels than any of the low-resolution images. However, in practice, these

shifts are randomly distributed due to the uncoordinated nature of the fractional pixel motions and the relative rotations of each frame in the sequence.

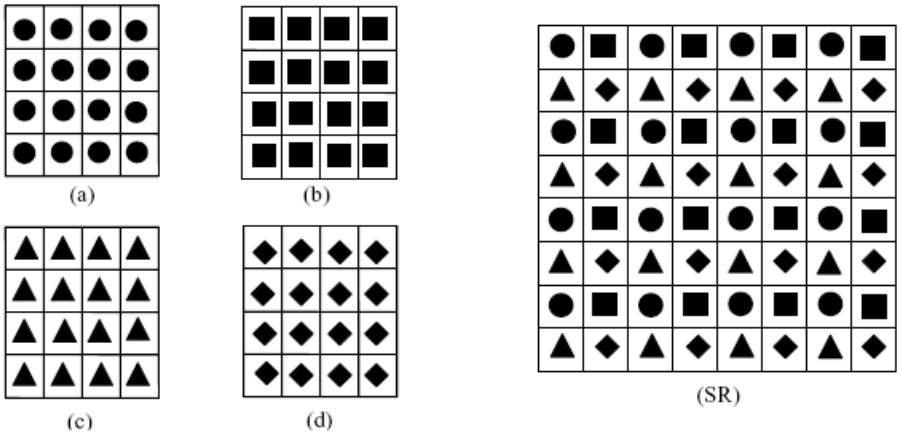


Figure 2 – An idealised Super Resolution (SR) set-up

Hence, these random motions or shifts must be known accurately in order to create a regular and refined grid of interpolated pixel values. Since the interpolation process is an estimation process which determines the pixel brightness which would exist on the intersections of a regular grid using randomly spaced pixel locations (representing the low-resolution images), several interpolators may be used depending on the application and accuracy requirements. One of the methods for interpolating scattered data to a uniform refined grid is referred to as *Universal Kriging*.

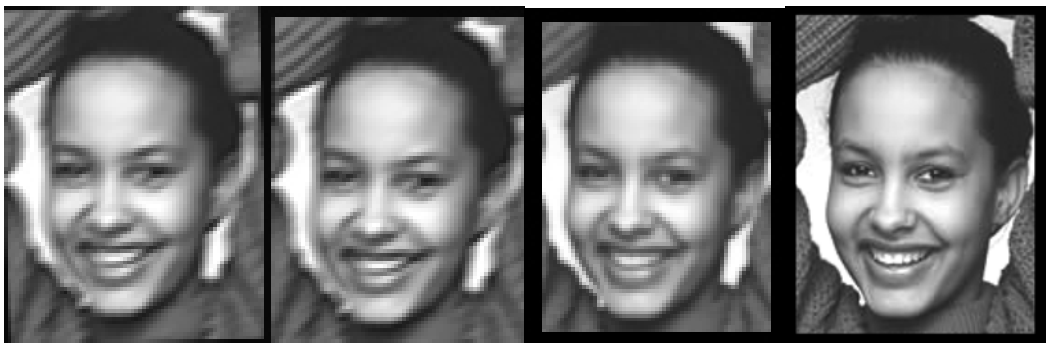
This method was used in the following tests because is a statistical interpolation technique that considers both the distance and the degree of variation between known data points when estimating values in unknown areas. A kriged estimate is a weighted linear combination of the known sample values around the point to be estimated. Kriging allows the user to derive weights that result in optimal and unbiased estimates. It attempts to minimize the error variance and set the mean of the prediction errors to zero so that there are no over- or under-estimates.

An important feature of Kriging, as compared with other image or surface interpolators, is that it gives an estimation of the error at each interpolated point, thus providing a measure of confidence in the modeled surface. A thorough theoretical explanation of Kriging interpolation is beyond the scope of this paper and the reader is referred to Rees (2007) for the theory and applications of this interpolation technique in the particular areas of digital imaging and remotely sensed data.

6 Tests and results

In order to test the efficacy of the video stabilisation software and the image enhancement process a set of 256 simulated images were generated. In this instance the ROI was the face shown in Figure 3 (70x90 pixels). The motion between the frames follows a generic affine model described in section 2 where the translation and small rotations between frames were randomly generated. The 256 images were stabilised using the *Deshaker* software.

Prior to stabilisation the average frame-to-frame displacement was 5 pixels with rotation angles not greater that 5^0. After stabilisation, the average pixel displacement was 0.5 pixels. Visually, little or no motion could be seen in the stabilised video. The combination via super-solution of all the stabilised images produced the improved results shown in Figure 3-d. The magnification factor used in the reconstruction of the face was 4.0. That is, the enhanced composite produced an image with dimension 280x360.



(a) (b) (c) (d)

Figure 3 – (a) and (b) are two examples (70x90 pixels) of the 256 de-stabilised video frames. (c) is one of the stabilised frames whereas (d) is the result of combining 256 stabilised frames via super-resolution (280x360 pixels).

In this synthetic test care was taken to minimize the effect of rotations in the process. Strong rotations in the images may have required additional cropping/trimming of the images with repercussions on the final enhancement. Thus detracting from the strength of the conclusions reached in the experiment. Correlation obviously exists amongst the video camera and orientation parameters such as tilts, rotations and affinity/obliquity of the sensor, and, in a controlled experiment where the aim is to demonstrate the use of a process to enhance image resolution *per se*, it was thought unwise to introduce such complications.

Although this experiment relates to a grey scale sequence, the same process can be applied when using colour. Colour images can be considered as three separate images containing red, green and blue components (RGB). Each of

these components or channels can be enhanced independently and then fused to produce an enhanced colour image with enhanced resolution.

For the same video settings and characteristics a second test was devised involving the scene shown in Figure 4. In this more realistic case the unstable video sequence involved more complex motions. The ROI is the license plate of the vehicle. The total length of the video was 60 seconds and 256 images were extracted, stabilised, combined and super-resolved so as to achieve the improved results shown in Figure 3(c). Similar to the previous test the magnification factor applied in the reconstruction was equal to 4.

(a) (b) (c)

Figure 4 - (a) is one of the video frames of the vehicle prior to stabilisation (180x280). (b) is one frame after applying the stabilisation, and (c) is the combination of all 256 frames as extracted from the stabilised video sequence (540x840).

Discussion and Conclusions

- Digital video stabilisation techniques have been studied for decades to improve visual quality of image sequences captured by compact and light weight digital video cameras. When such cameras are hand held or mounted on unstable platforms, the captured video generally appear shaky.

- Unwanted video vibrations would lead to degraded views which would also greatly affect the performance of applications such as video encoding and video surveillance. With recent advances in wireless technology, video stabilisation systems are also considered for integration into wireless video communication equipments for stabilisation of acquired sequences before transmission, not only to improve visual quality but also to increase the compression performance.

- The problem of video resolution enhancement can be addressed by exploiting multiple stabilised frames 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 video capturing device. However, exploiting these motions requires accurate estimates of them.

- A method for extracting a higher quality still image 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, that is, video stabilisation and image super-resolution. The method does not rely on control points for the process of matching or registering the images.

- The application and effectiveness of the enhancement process has been demonstrated in synthetic and real case scenarios. Refinements to the technique are being undertaken to decrease the processing time and increase the accuracy achievable for larger image magnifications. 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.

References

Z. Wang, H. R. Sheikh, and A. C. Bovik, 2003 Objective video quality assessment," in *The Handbook of Video Databases: Design and Applications*, B. Furth and O.Marqure, Eds., chapter 41, pp. 1041–1078. CRC Press.

L. Marcenaro, G. Vernazza, and C.S. Regazzoni, 2001.Image stabilisation algorithms for video-surveillance applications" in *International Conference on Image Processing*, pp. I: 349–352.

Rees W. G. 2007. "Physical Principles of Remote Sensing". Second edition. Cambridge University Press.

Gonzalez R.C. and Woods R.E. 2007. Digital image processing. 3d. Edition, Published by Prentice Hall. 954 pages.

C. Morimoto and R. Chellappa, 1998 "Evaluation of image stabilisation algorithms. in Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, vol. 5, pp. 2789-2792.

Zhouchen, L. and Heung-Yeung, S., 2004. Fundamental limits of Reconstruction-Based SR Algorithms under Local Translation. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 26, No.1, January.

M. Tico, S. Alenius, M. Vehvilainen, 2006 Method of Motion Estimation for Image Stabilisation, in Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 277-280.

C. Tomasi and T. Kanade, 1991. Detection and Tracking of Point Features. Carnegie Mellon Univ., Tech. Report. CMU-CS-91-132.

A.U. Batur and B. Flinchbaugh, 2006 . "Video Stabilisation with Optimized Motion Estimation Resolution", Proc. of International Conference on Image Processing, pp.465-468.

Bovik A. C. 2009. The essential guide to video processing. Academic Press. 755 pages.

S. Auberger, C. Miro, 2005 "Digital Video Stabilisation Architecture for Low Cost Devices", in Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, pp. 266-271, September.

Y. Matsushita, E. Ofek, G. Weina, T. Xiaoou, S. Heung-Yeung, 2006 "Full-frame Video Stabilisation with Motion Inpainting", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 7, pp. 1150-1163, July.

S. Battiato, G. Gallo, G. Puglisi, S. Scellato, 2007 "SIFT Features Tracking for Video Stabilisation," in Proceeding of the IEEE International Conference on Image Analysis and Application, pp. 825–830, September.

Z. Wang, H. R. Sheikh, and A. C. Bovik, 2003 "Objective video-quality assessment," in *The Handbook of Video Databases: Design and Applications*, B. Furth and O.Marqure, Eds., chapter 41, pp. 1041–1078. CRC Press.

Pilgrim L.J. 1991. "Simultaneous Three Dimensional Object Matching and Surface Difference Detection in a Minimally Restrained Environment". PhD Thesis No. 066.08.1991. Department of Civil., Surveying and Environmental Engineering. The University of Newcastle, Australia. 215 pages.

Farsiu S., Robinson D., Elad M., and Milanfar P. 2004 "Advances and Challenges in Super-Resolution", International Journal of Imaging Systems and Technology, Volume 14, no 2, pp. 47-57, August.

Yu-Ming Liang, Hsiao-Rong Tyan, Shyang-Lih Chang, Hong-Yuan Mark Liao, and Sei-Wang Chen. 2004. "Video Stabilisation for a Camcorder Mounted on a

Moving Vehicle". IEEE Transactions on vehicular Technology, Vol. 53, No. 6. November.

Wolf P. and DeWitt B. 2000. "Elements of Photogrammetry with Applications in GIS". McGraw-Hill Science/Engineering/Math; 3d. edition. 624 pages