Hierarchical Block-based Image Registration for Computing Multiple Image Motions

Gareth van Essen
School of Engineering and
Advanced Technology
Massey University
Palmerston North, New Zealand
Email: gwvanessen@gmail.com

Stephen Marsland
School of Engineering and
Advanced Technology
Massey University
Palmerston North, New Zealand
Email: s.r.marsland@massey.ac.nz

John Lewis
Weta Digital and
Massey University
Wellington, New Zealand
Email: zilla@computer.org

***Abstract*—This paper proposes a block-based approach to the problem of image registration for detecting camera motion in the presence of moving objects, intended for application in the area of video inpainting for the film industry. The method of aggregating search results on individual blocks during the registration offers a simple and effective way to isolate the background transform, as well as offering approximate segmentation of the moving objects in the scene without extra computing overhead.**

I. Introduction

The purpose of this paper is to outline a technique for finding the registration between two frames from a sequence of video that corresponds to the camera motion, that also provides a means of detecting, and approximately segmenting, the moving objects within that scene. The approach taken is to split the image up into a series of blocks, and run a search comparing different registration parameters to find the transform that most effectively maps the motion of the scene background by means of grouping.

A sequence of video generally involves two main unknowns: the movement of the camera, and the form and motion of the moving objects in the scene. This can involve many different motions, all of which must be detected and calculated seperately, which is a very difficult task.

Image registration is the process of matching a pair of similar images in terms of the rotation, translation, scale and shear required to make those images correctly align. The particular focus of this paper is the application of this on two frames from a video sequence to describe image motion. Applications include automated image and video inpainting for special effects in the film industry, use with mobile security systems, and vision systems for autonomous agents.

For images that do not contain moving objects, the process of extracting information about the movement of the camera for the scene is relatively straightforward – by using an affine transformation model it is simply a matter of finding transformation parameters (rotation and translation) that match the two images using a simple error metric like sum of squares difference. However, once moving objects are introduced to the scene, the registration process becomes more complicated. For a situation where the motion of the camera is unknown, and the location of any moving objects is unknown, a registration algorithm has a significant amount of information that it needs to infer.

Different sections of the images will have different motions (for example the background moving in one direction due to camera motion, while a person walking through the scene moves in a different direction). These areas of separate motion need to be identified and the magnitude and direction of the motion needs to be independently calculated.

This paper looks at an approach to this problem using a block-based technique for the registration. This enables differentiation between the dominant object in the scene (assumed to be the background) and other objects, which move relative to the dominant object. This provides information on the motion of the camera filming the scene.

The task of identifying moving objects within a scene is another area of computer vision that is attracting attention. This paper proposes to use the information obtained through the process of registering the two images to provide a reasonable identification of the object location, for eventual application in the area of video inpainting.

II. Method

A. *Affine Registration*

The technique used for registration in this paper is parametric affine registration. We currently restrict the algorithm from full affine to a three-parameter model, using rotation and x- and y- translation. The decision to ignore shear and scale considers that the effect of these between adjacent frames in standard video is likely to be minimal, and the block-based approach should allow for this to be detected without explicitly searching for it. Though it may have some impact on accuracy, the gain in efficiency obtained by ignoring scale and shear is significant.

The three-parameter affine transformation model for mapping the registration of a single pixel between two images (I_0 and I_1) is as follows

$$f(x_0, y_0) = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$

978-1-4244-4698-8/09/$25.00 ©2009 IEEE

where $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ is the position of a pixel in the original image I_0 and $\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$ is the position of the corresponding pixel in the second image, I_1. The rotation is conducted about the point $(0, 0)$ on the image (the top left corner), so for a different centre of rotation the image must be offset. The values t_x and t_y are the x- and y-translations respectively.

The registration itself is a minimisation problem: identifying values for t_x, t_y, and θ that minimise an error value between image I_1 and the transformed I_0. The error used for this paper is the sum-of-squares difference between the RGB colour values of each pixel over the image or image section:

$$\epsilon = \sum_{i,j} (I_0(x_i, y_j) - I_1(f(x_i, y_j)))^2$$

As algorithmic efficiency is a major concern, the use of a gradient descent algorithm to optimise the search was investigated. For this application, however, gradient descent is not ideal – the frequent occurrence of local minima throughout the search space, a high susceptibility to noise, and an inability to effectively process areas of low colour differential severely hamper its effectiveness.

The efficiency problems can be greatly reduced with an approach using a series of iterative searches over reduced-scale images. The scale of the reduction required depends on the image sizes involved – for the relatively small samples analysed in this project, an initial coarse search is conducted over a quarter-scale image, which is then refined over a half-scale image before a full-scale image is used to obtain the sub-pixel accuracy required.

B. The Block-based Approach

Attempting to match a transform to the entirety of an image is impractical for a registration technique on video containing moving objects, as different parts of the image will be moving different amounts. Therefore, it makes sense to acknowledge that there will be objects, and allow them to be excluded from the registration. Methods like optical flow can achieve this by mapping pixels individually, but individual pixels are highly susceptible to noise, which can in turn affect the registration.

If working over the entire image is too broad, and looking at individual pixels is too fine, then it is logical to try and find a sort of middle ground. By splitting the image up into a series of blocks and tracking these separately, errors caused by certain types of noise (most notably, the blurring that is a common result of video compression) can be minimised, and sections of the image containing moving objects can be eliminated from the registration problem.

There are two approaches to the block-based image registration. The first involves running the registration search algorithm for each block individually to calculate the best transform for each block. These results can then be correlated in the form of a three dimensional histogram, where the largest group corresponds to the transformation that describes the motion of the largest ‘object’, which can be taken to be the background.

This method is inherently slow, as the search space is fairly large. For this reason, an iterative search approach over a multi-resolution pyramid is employed. The algorithm is initially run at quarter-scale or smaller (depending on the size of the original image) with broad search parameters, then the results are aggregated and used to determine a much smaller search area, which is then applied to a half-scale image. The search parameters are then refined again, before finally being applied to the original full-scale image.

Susceptibility to noise and difficulty distinguishing large areas of colour lead to a proportion of blocks returning incorrect results for the registration. For this reason, at each stage of the refinement the overall best result is used as the base for the next iteration of the registration for each block.

The second method is considerably faster, but also introduces a much greater potential for error. It involves picking a block at random and running the affine search over that sole block to find the transformation that fits that block to the next image. That transformation is then applied to all the blocks and thresholding is used to identify which blocks agree with the transform, which in turn gives the approximate area of the object whose movement is mapped by the transform.

Largely, this causes problems because the system is based on arbitrary thresholds. If the threshold identifying which blocks agree with a transform is too low, blocks which do agree may be excluded, resulting in a lower value for object area, which could result in the background being disregarded as being just a small object. If the threshold is too high, the algorithm may find a local minima that is deemed an acceptable transform, but is not the global minima – most likely to occur when the sampled block contains parts of more than one object.

The other threshold that may cause problems is the one that is used to determine whether an object is large enough to be identified as the background. If the value is too high then there may be no objects that meet the criteria and the algorithm will never complete, or, conversely, if it is set too low then a smaller object may be incorrectly identified as the background. This particular problem may be countered by eliminating blocks that have already been grouped and resampling until every region of the image has been identified, at which point the largest of these can be positively identified as the background. This will, however, significantly increase the computational cost of the algorithm (as it will no longer halt as soon as it finds an acceptably large object) as well as introducing the possibility of error when a block is assigned into a group that ‘fits’ but is not the optimum grouping for the block.

In theory, this algorithm appears to solve the problem that is the target of this paper: identifying and tracking each individual object in the scene. In practice, however, this is unachievable. Firstly, there are the blocks that contain parts of more than one object; these will seldom match the transform of any object, so will not be grouped correctly. Similarly, blocks that are resampled to by the algorithm and that exist in areas

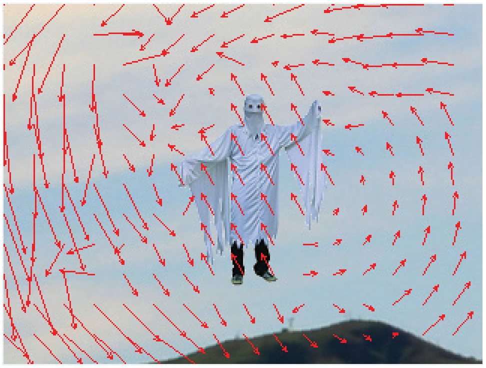

Fig. 1. This shows the detected motion paths of blocks across the image after one iteration of the affine search on the full-scale image. (see Figure 3 for the images used in this registration). The necessity of an iterative search approach is shown by the number of apparently random motion paths that are produced.

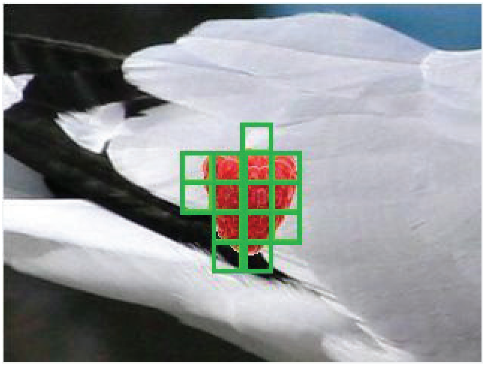

Fig. 2. Example of detection of a moving object using the block-based method. The image was first registered to a second image, then the block information was used to highlight moving objects within the scene. The blocks shown on the image indicate the detected object area.

of solid colour (i.e. have no distinguishing features/colours) will likely not track correctly, which would also interfere with the results.

Figure 1 shows the chosen transformations for a registration search run over an image pair using 20x20 blocks. Only a single search is run on the full-scale image, to demonstrate the registration process without the corrective measures that are used in the iterative coarse-to-fine search. The number of blocks that produce conflicting registrations demonstrates the high occurrence of local minima over the search, and thus the necessity of the corrective groupings used in the iterative search.

C. *Extracting Object Information*

The extraction of object information is based on the premise that most of the image is background, which moves according to the motion of the camera, with certain sections of the image not conforming to this motion – those sections belonging to independently moving objects.

During the process of tracking the camera motion, the image sections that do not conform to the global motion must be detected and eliminated in order for the result to be accurate. Thus, the location and, to a certain extent, the motion of these objects is already available once the camera tracking algorithm has run

Due to the necessarily coarse nature of the block-based approach, an exact outline of the objects is not provided (see Figure 2), however enough information is discovered to make the task of accurately segmenting these objects much easier with the application of an algorithm such as active contours.

For the purpose of comparison, a second method based on a difference algorithm applied post-registration is used. The algorithm uses the existing block structure to reduce the impact of noise on the results. Each block is marked as belonging to an independently moving object if sufficient pixels within the block are significantly different. The level of difference that is taken as significant is determined by an adjustable threshold, which requires changes depending on the level of contrast between background and foreground in the samples.

The method using registration information to identify moving objects works well in a broader range of situations. While the difference algorithm is often slightly more effective in high contrast scenarios, it struggles with low contrast.

III. Experimental Results

A. *Object Detection*

For this section, the algorithm developed in this paper, which builds object information as a consequence of the registration, is compared against a difference-based object detection method that is applied post-registration.

Figure 3 shows a pair of images that will be used for this test. The images have a moving object, and the camera is shifted and rotated between frames. Block size of 20×20 is used.

Upon application of the first object detection method, the result is shown in Figure 4. The image displayed is the mean of the images once they have been registered and aligned, allowing the two locations of the object to be visible simultaneously. The blocks identified as not belonging to the background give a fair approximation of the object. Figure 5 shows the same sequence processed by the difference-based method. The results are not massively different, with the registration-based detection giving slightly better coverage of the object. The main difference is that the difference-based method required a significant amount of threshold tweaking to get to this point, where the registration method did not.

B. *Block Size*

The block sizes tested for the current incarnation of the algorithm are all square, mainly to keep things simple. The use of rectangular and other shaped blocks at this stage is unlikely to have any benefit, however this may be explored for the multi-block extension detailed in Section V.

The block sizes looked at for the test sequences (which measure 320×240 pixels) are as follows: $5 \times 5, 10 \times 10, 15 \times 15, 20 \times 20, 25 \times 25, 30 \times 30$, and 40×40. The same sized block is applied to the reduced-scale images used for the initial

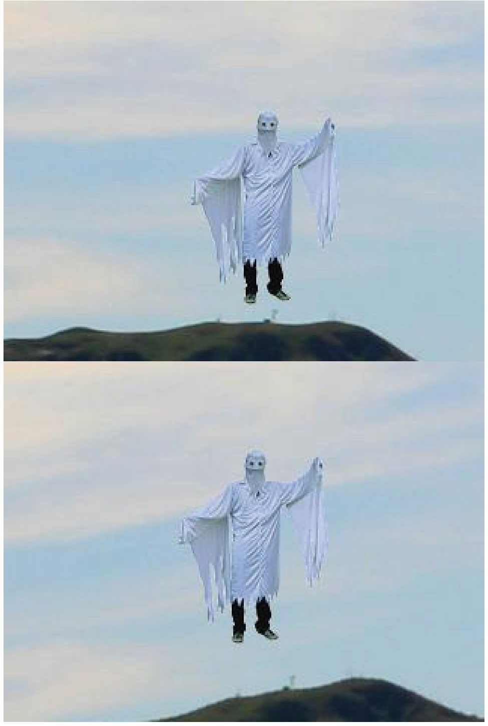

Fig. 3. Sample image pair for registration involving camera motion and an independently moving object.

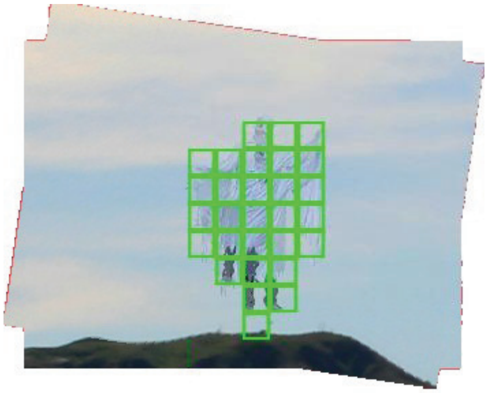

Fig. 4. Object detection using results of the block-based registration

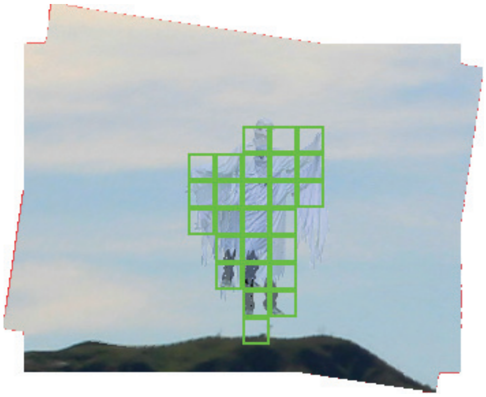

Fig. 5. Object detection using difference-based method

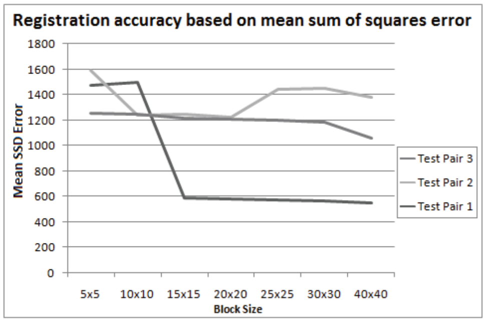

Fig. 6. This shows the accuracy of the registrations on different test images using different block sizes. Three image pairs were used, based on images with different image statisctics. Error values are mean sum-of-squares difference.

coarse search conducted by the full-spread search algorithm (see Section II-B), as the minimum number of pixels required to produce accurate results does not change as the image gets smaller, since the size of a pixel remains constant. Figure 6 shows the mean sum of squares error for registration with different block sizes over three different image pairs.

When the algorithm is run with block sizes of 5×5 and 10×10, the number of pixels in each block is insufficient to accurately distinguish unique sections of the image and correctly map them in the registration. As a result, the registration at these block sizes is not sufficiently accurate to be useful. The object detection at this level is also not very effective, as essentially anywhere in the image with significant edges are picked up as deviating from the overall transform (since the transform is incorrect).

With a block-size of 15×15 to 25×25 the registration result remains the same for most image sequences, but in some cases 20×20 will give a more accurate result. In this case it can become a trade-off situation, since the smaller block size will give better object detail, but in most cases registration accuracy is the crucial element so a block size of 20×20 is best.

Beyond a block size of 30×30, the graph indicates that the registration is effective, however this is a limitation of using sum-of-squares difference as a measure of error when there are multiple objects in the image – the actual registration result for two of the test pairs is very poor, but the error value is low. The problem with a larger block size is that the proportion of blocks that contain multiple objects is high, which interferes with the registration.

Another consideration for the algorithm is the capability for dealing with noise. This capability is dependent on the type of noise present: if the noise is in the form of a small number of clustered noise pixels across the image, these clusters will simply be treated as separate objects and captured by the algorithm; with speckle noise (see Figure 7), however, the entire registration can be affected. Figure 8 shows the accuracy of the registration for images containing different levels of speckle noise using different block sizes. This indicates that larger block sizes enable the algorithm to more easily cope with high levels of noise.

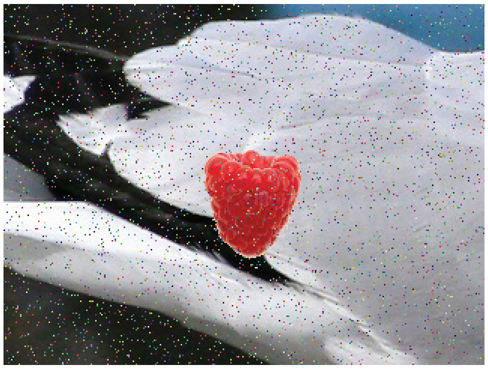

Fig. 7. Example image with 5% speckle noise. The amount of damage is substantial.

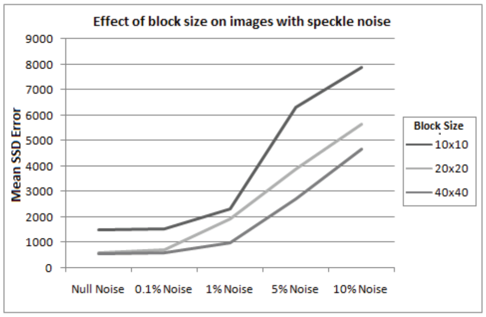

Fig. 8. Different amounts of speckle noise applied to sample images. This graph shows the accuracy of the registration using different block sizes for different noise levels. The difference in accuracy between the block sizes does not change very much over the changing noise levels, and beyond about 1% noise the accuracy is degraded too much to be useful.

IV. Related Work

There are many methods of affine registration; good over–views of the work in the area are offered by [1], [2]. A common application makes use of registration techniques for mapping medical imaging scans to assist in the detection of anomalies, however the types of deformation involved do not transfer to the area of camera motion effectively.

Patch-tracking is one area within the scope of registration and object tracking that has been pursued, with two different interpretations of its meaning available. The patch-tracking algorithm put forward in [3] is focused on tracking an object, and requires prior knowledge about the form of that object, which is not particularly useful here. [4] offers an entirely different patch-based algorithm, which uses patches similar to the blocks used in this paper to solve regression equations for image distortion. The results presented here are impressive, however it does not deal with the existence of moving objects in the scene.

Window-based algorithms offer a similar sort of idea under a different name, such as those described in [5], [6]. These algorithms are designed to identify the camera motion in scenes with moving objects, and make use of gradient descent to find both the global minimum and a secondary, local minimum that describes the camera motion – this is specifically targeting scenes that are *not* dominated by background.

In terms of a registration approach to object segmentation, [7] establishes an effective foundation for work in this area, but the motion results contain some sections of image unrelated to the objects. [8] presents with a more complete example of this approach, which is very effective, but the results are still influenced by outliers in the sequence. [9] also provides some work in this area, but the resulting object segmentation displayed in the paper is patchy and incomplete. Another approach to the problem, using a pixel-wise optical flow method for registering and segmenting an image into its components, is presented in [10]. It produces some impressive results, but due to its pixel-wise nature the layers of the image will not separate completely, with some pixels being assigned incorrectly.

The work presented in [7] is somewhat in between. A block based approach is used for the object detection stage but not for the global motion estimation, and the object detection uses a difference based approach rather than using the information available from the registration.

V. Multi-block Mapping

By introducing extra blocks across the image that overlap but are offset from the main registration blocks before the final fine-search of the registration, a higher degree of accuracy can be achieved for the object tracking at a certain cost in terms of computing power. These extra blocks are registered in the same way as the original blocks, and the results are aggregated using the logical AND operator to produce a finer resolution for the tracking result – only areas that are marked as moving objects by *all* of the blocks that contain them are included in the tracking result. This can be seen as a basic form of image super-resolution.

For example, using one extra set of pixel blocks, offset from the original grid by 10 pixels in both the x- and y-directions, would produce a tracking result constructed of 5×5 pixel blocks rather than the 20×20 blocks that would otherwise be the result.

Figure 9 shows both the basic approach and the multi-block approach (using three extra block sets over the image). The multi-block approach provides a segmentation that does not include as much non-object data, but it still loses small sections of the edge of the object.

VI. Conclusions

This paper offers a method for using a block-based approach for identifying the motion of the background between two frames of a background-dominant video sequence. The preliminary results indicate that the accuracy of this registration is good.

Also presented here is a method for identifying moving objects within a scene, without the need for further processing once the registration process has been completed. As shown in the results, the accuracy of this segmentation requires further work, but the results that have been obtained at this stage show that the method is likely to have potential applications, particularly for area of video inpainting, which is to be the

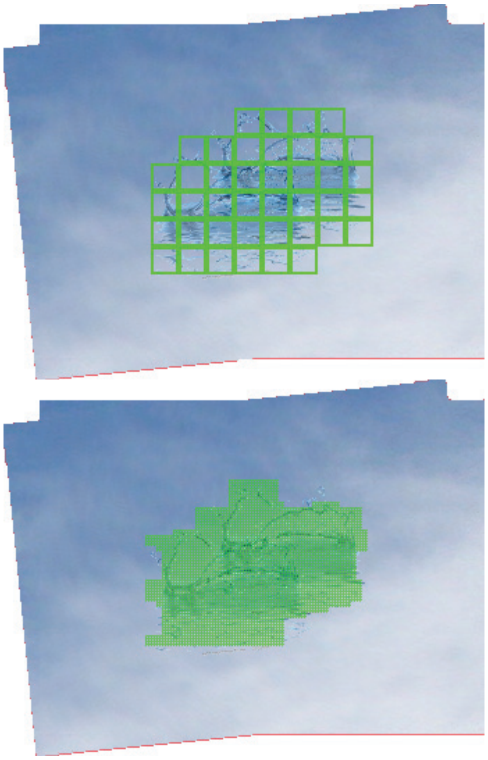

Fig. 9. The first image shows the object detection using the basic approach, with the detected object area shown by the registration blocks it is detected in. The second image shows the multi-block approach, this time the detected area is shown as a grid to simplify the drawing algorithm. The images that the object detection is shown on is the mean image taken after registration (allowing the position of the object in both frames to be seen on the same picture).

future direction of this project. Since the frame backgrounds have been registered, sections of different frames can easily be transferred from one frame to the other. Also, the sections that belong to the background are known, so the process of inpainting for sections where the background is unobscured at some point during the sequence will be fairly simple.

VII. Future Work

The direction of the project in the future focuses on the adaptation of this method for two-frame registration and object detection for use with full sequences of video. This process should be relatively simple, and the fact that motion between temporally close frames are not independent should result in a reasonably low computational cost.

The first part of the algorithm – the process of searching for an appropriate registration with the block based approach – should extend fairly easily for use with video. The registration can be determined between neighbouring frames in the sequence, giving sufficient transformation information to extract a match between any two frames in the sequence.

As demonstrated in this paper, the process of determining the registrations will provide some information about the moving objects in the scene. The accuracy of this information should then be able to be enhanced beyond what was possible in the two image model, by aggregating the information between multiple nearby frames. In particular, this offers the opportunity to identify which sections of a detected motion are background and which actually belong to a moving object.

The next step beyond this is to make use of the registration and object information to produce a method for automated inpainting of video sequences.

The level of object detection provided by the algorithms described in this paper provide a very effective platform on which to deploy a segmentation algorithm such as those described in [11]. As a basic outline of the object(s) is already provided, the segmentation will be both faster and more accurate than if it were run on the overall frame.

References

[1] M. Shah and R. Kumar (Eds). *Video Registration*. Kluwer, May 2003.

[2] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, October 2003.

[3] Jose Miguel Buenaposada, Enrique Munoz, and Luis Baumela. Tracking a planar patch by additive image registration. 2003.

[4] P. Zhilkin and M. E. Alexander. A patch algorithm for fast registration of distortions. *Vibrational Spectroscopy*, 28(1):67 – 72, 2002.

[5] A. Krutz, M. Frater, M. Kunter, and T. Sikora. Windowed image registration for robust mosaicing of scenes with large background occlusions. In *Image Processing, 2006 IEEE International Conference on*, pages 353–356, Oct. 2006.

[6] A. Krutz, M. Frater, and T. Sikora. Window-based image registration using variable window sizes. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 5, pages V –369–V –372, 16 2007-Oct. 19 2007.

[7] Bin Qi, M. Ghazal, and A. Amer. Robust global motion estimation oriented to video object segmentation. *Image Processing, IEEE Transactions on*, 17(6):958–967, June 2008.

[8] Marina Georgia Arvanitidou, Alexander Glantz, Andreas Krutz, Thomas Sikora, Marta Mrak, and Ahmet Kondoz. Global motion estimation using variable block sizes and its application to object segmentation. *Image Analysis for Multimedia Interactive Services, International Workshop on*, pages 173–176, 2009.

[9] Fang Zhu, Ping Xue, and Eeping Ong. Low-complexity global motion estimation based on content analysis. In *Circuits and Systems, 2003. ISCAS '03. Proceedings of the 2003 International Symposium on*, volume 2, pages II–624–II–627 vol.2, May 2003.

[10] John Y. A. Wang and Edward H. Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing*, 3:625–638, 1994.

[11] Tony F. Chan and Jianhong Shen. *Image Processing and Analysis*. SIAM, 2005.