STITCHING LARGE MAPS FROM VIDEOS TAKEN BY A CAMERA MOVING CLOSE OVER A PLANE USING HOMOGRAPHY DECOMPOSITION

E. Michaelsen

Fraunhofer IOSB, Gutleuthausstrasse 1, 76275 Ettlingen, Germany
eckart.michaelsen@iosb.fraunhofer.de

KEY WORDS: Panorama stitching, planar homographies, homography decomposition, underwater mapping, UAV surveillance

ABSTRACT:

For applications such as underwater monitoring a platform with a camera will be moving close to a large roughly planar scene. The idea to map the scene by stitching a panorama using planar homographies is nearby. However, serious problems occur with drift caused by uncertainty in the estimation of the matrices and un-modelled lens distortions. Sooner or later image points will be mapped to infinity. Instead this contribution recommends using the homographies only for the composition of local patches. Then the homography obtained between the first and the last frame in such patch can be decomposed giving an estimate of the surface normal. Thus the patches can be rectified and finally stitched into a global panorama using only shift in *x* and *y*. The paper reports about experiments carried out preliminarily with a video taken on dry ground but a first under water video has also been processed.

1. INTRODUCTION

1.1 Intended Applications

In particular underwater robot vision is restricted to keep the distance between a structure to be monitored and the platform on which the camera is mounted short. There are ideas to enlarge the allowable distance by using gated viewing devices [9], but in the waters found where the application is supposed to be located there will always be a maximal distance where no considerable image quality is allowed due to floating obfuscation. Good image quality can often be expected from imagery taken at distances such as one meter. On the other hand the structure to be mapped may well have an extension of several hundred meters. Here we restrict ourselves to roughly and locally planar structures – such as retaining walls, harbour structures or underwater biotopes. The goal is to stitch a kind of orthophoto from a long video sequence.
Under water the drift problem – as outlined in section 1.2 - is very serious. But it also occurs in unmanned aerial vehicle mapping, where the platform may be cruising in about a hundred meter height over a roughly and locally planar world of much larger extension, e.g. several kilometres in extension.

1.2 Problem

The standard state-of-the-art method for stitching of an image sequence into a larger panorama is driven by successive planar homography estimation from image to image correspondences between interest points. Most often it is assumed tacitly or explicitly that the camera should only rotate round the input pupil and not move around in space. If the scene is strictly planar, there is – in principle - no difference between the image obtained by a wide-angle view from close up (in pin-hole geometry and taken normally) and a view from further away, or even an ortho-normal map. So the stitching of large views using homographies should be equivalent to taking an ortho-normal map.
However, deviations from the planar scene form, e.g. when a retaining wall is only locally planar but cylindrical in its global shape, cannot be treated this way. Moreover, if the first frame of the video sets the reference – as is often done – it may well not be exactly normal to the scene. Then there exists a distance in which the plane through the camera location and normal to its focal axis will intersect the scene plane at a line somewhere. Points on this line will be mapped to infinity if the homography estimation were precise – and points beyond this line would appear on the opposite end of the panorama. If we are only one meter away from a structure of hundreds of meters this is to be expected.
More seriously, the homography sequence approach accumulates the inevitable errors in large chains of matrix multiplications. Such drift may contain un-biased parts from uncertainty in the interest point locations, but it also may contain biased parts. E.g. homography estimation tends to hide un-modelled lens distortions in the rotational part of the homography [2].

1.3 Related Work

Many panorama stitching software packages are commercially available or can be downloaded for free from the web such as HUGIN [1]. The theory of optimal estimation of homogenous entities, such as planar homography matrices, with more entries than degrees of freedom from image to image correspondences with proper uncertainty propagation has reached a high level of sophistication [5]. RANSAC methods for robust estimation of such entities are standard today [4,7] but there are also alternatives such as iterative reweighting or GoodSAC [11]. Under water panorama stitching has been addressed e.g. by [2] with particular emphasis on the lens distortion induced drift.

2. STITCHING LOCAL PANNOS INTO A LARGE MAP

2.1 Homography Estimation

A planar homography is a mapping $x'=h(x)$ from one image into the other keeping straight lines straight. Here x and x' respectively are the points in the images. Homographies form an

algebraic group with the identity as one-element. Using homogenous coordinates the homographies turn out to be linear: $x'=Hx$. Where H is a 3x3 matrix whose entries depend on the choice of the coordinate system in the images. This linear description hides the highly non-linear nature of homographies in the division when transforming x' back into inhomgenous image coordinates. Thus homographies may map a finite point into infinity and they are not invariant for statistically important entities such as centre of gravity or normal distributions. Still, there is consensus today that homographies can be estimated from a set of four or more correspondences of interest points using the linear matrix equation provided that 1) a coordinate system is used that balances the entries into the equation system such that signs have equal frequencies and absolutes are close to unity [7], and 2) H is not too far away from the unity matrix (in particular the "projective" entries H_{31} and H_{32} should be small). In a video sequence 1) can be forced and 2) can be assumed. Thus, we follow the usual procedure using an interest points, correlation for correspondence inspection, and RANSAC [4] as robust estimator. The activity diagram in figure 1) gives the details of the procedure. In each frame of the video a set of interest points $\{p_{in}\ ;\ i=1,\ \dots\ ,\ k_n\}$ is extracted using the well-known squared averaged gradient operator in its Köthe variant [6,8]. These are tracked back in the previous frame also using standard functions – here optical flow including image pyramids from open CV base [12]. Among these a consensus set is selected and simultaneously an optimal homagraphy using linear estimation and RANSAC on the correspondences of the p_{in} in coordinates transformed accordingly [7].

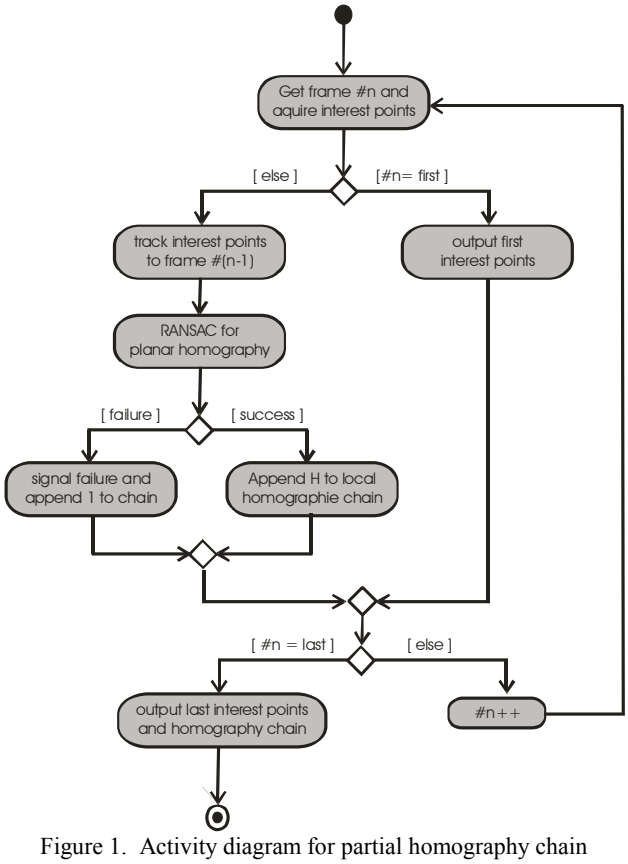

Figure 1. Activity diagram for partial homography chain estimation

Homographies cannot only be estimated for successive video frames but also for frames further apart from each other as long as there is sufficient overlap. However, if there is no sufficient overlap anymore the homographies must be chained in a sequence – by successive multiplication of matrices. Since there is uncertainty in the entries of this product there will be a drift – also in the "projective" entries H_{31} and H_{32}. Sooner or later points from an image far away from the first frame will thus be mapped to infinity.

2.2 Homography Decomposition and Rectification

Here H must be given in the normalized form, i.e. with the image coordinate system transformed such that the focal length equals unity and the principle point of the camera equals the origin of the coordinate system. So focal length and principle point should be known in good approximation. The standard decomposition of the matrix H in the form

$$H = \lambda R + tn^T \tag{1}$$

is known since [3]. Here R is the rotation matrix of the camera between the images, t is a translation vector, n is the surface normal of the planar scene, and λ a scalar factor. t can also be interpreted as homogenous entity. Than it is the image of the other camera, the epipole. n can also be interpreted as homogenous line equation. Than it is the line at infinity or the horizon of the scene.
This is the most important result here. The application demands that a proper – close to orthonormal – mapping of the scene should yield $n_1=n_2=0$ i.e. the normal identical to the viewing direction. After decomposition of H this can be achieved by applying appropriate rotations round the x and y axes:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix} \text{ and } \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix} \tag{2}$$

where $\beta=atan(n_1/n_3)$ and $\alpha=atan(n_2/n_3)$ after the rotation round the y axis. With this transformation the view should be rectified. We refer to [10] for a detailed analysis of the decomposition. There also a purely analytical solution to the decomposition can be found using only roots. Here the classical singular value decomposition version is used decomposing H into a product $H=UDV$. The entries of the central diagonal matrix d_{11}, d_{22}, d_{33} are the critical parts. They must be of sufficiently different sizes. Their differences are used as denominator while solving the quadratic equation system.
Two significantly different solutions appear among which we pick the one with n closest to $(0,0,1)^T$. The other solutions are flipped sign versions of no interest. But if the two solutions are equally close to $(0,0,1)^T$ or if the singular values are too similar the decomposition fails (resulting in a failure branch in the flow in Figure 2).

2.3 Stitching the Local Patches into a Large Panorama

It is our intention to treat all rectified panorama patches equal. Neither any projective distortion should be applied to them anymore – since this was corrected by homography estimation, decomposition and rectification, nor any shearing – since this is excluded by sensor construction, nor any scaling – since we assume that the platform is capable of sensing, controlling and keeping its distance to the scene plane. The rotations round the x and y axes where fixed in the rectification step. We will also assume that the camera is not rotating round the z axis on the long run by means of appropriate other sensors on-board the

platform – e.g. gravity sensor under water or compass on a UAV. The only two remaining degrees of freedom are shift in *x* and *y* direction.

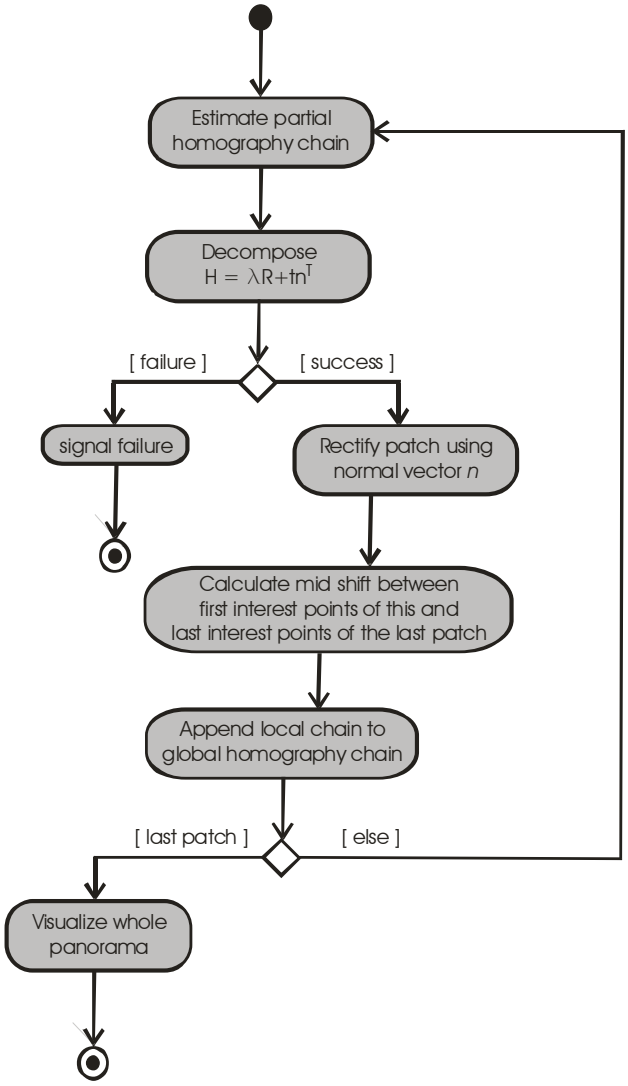

Figure 2. Activity diagram for global homography chain composition

This translation can easily be obtained by averaging the shift between the $p_{i,last}$ and $p_{i,first}$ of two successive patches. Recall, that the first image of a new patch is identical with the last image of the previous patch. Running the interest operator with the same parameters on the same image will give the same number of interest points in the same sequence. Such algorithms are deterministic. $p_{i,last}$ and $p_{i,first}$ of two successive patches are subject to different homographies, $p_{i,last}$ as result of a chained homography estimation plus a rectification and $p_{i,lfirst}$ as result of only a rectification. So there will be a residuum in this averaging process, which quantifies the success of the approach. But there cannot be any outliers.

Again a UML activity diagram gives an overview over this procedure (Figure 2). Here the stitching of a local panorama patch – i.e. the estimation of a partial homography chain as given in Figure 1 is hidden in one node.
This is still dead reckoning – since there is a possibly long sum of successive vectors with uncertainty drift – but it is much more stable than the multiplicative drift of the matrix chain. It is impossible that image points will ever be mapped to infinity

2.4 Resampling a Panorama from a Video

The main output of both a local estimation for patches as well as the global estimation for a panorama is a chain of homographies. So for each frame *i* of the video there is a homography h_i mapping a location – i.e. a line- and column index - of the panorama $(l_p , c_p)^T$ to a location in the *i*-th video frame $(l_{vi} , c_{vi})^T = h_i(l_p , c_p)^T$. However, a homography is a function mapping continuous coordinates into continuous coordinates. So if the panorama has similar or higher resolution than the video some type of interpolation will be required in order to fill the panorama with gray-values or colours from the video. Here the panorama is usually of lower resolution. So the coordinates in the video frame can be obtained simply by rounding $(l_{vi} , c_{vi})^T$.

Moreover, several frames of the video may contribute to the gray-value or colour to be displayed in one panorama pixel. The following possibilities are discussed:

- Averaging the value from all accessible frame locations $\{(l_{vi} , c_{vi})^T; 1 \le l_{vi} \le l_{max} \text{ and } 1 \le c_{vi} \le c_{max}\}$. This treats all information equally, but may give fuzzy results.
- Maximizing the gray-value over the index *i*. This is fast and easy, because all the non accessible positions either yield zero or NAN, but it has a bias towards brighter areas.
- Minimizing the distance to the centre $(l_c , c_c)^T$ using any metric $d_i = d((l_{vi} , c_{vi})^T , (l_c , c_c)^T)$ picks the gray-value from one particular frame. Here faults in the estimation may show up as sharp edges. We used this option here in order to explicitly show such problems.
- Maximizing the probability of a gray-value or colour given a drift model for the homographies and the measurements in the images (l_{vi} , c_{vi}) [13]. This needs assumptions on the uncertainty (e.g. normal distribution) and estimation of the parameters. Essentially, it leads to weighted averaging giving higher weight to gray-values from the centre. This needs most computational effort and diligence in parameter estimation – but leads to best and seamless results.

3. EXPERIMENTS

Some experiments where done outside of the water with a video taken by an Olympus PEN E-P1 camera with the standard zoom lens set to the extreme wide angle f=14mm. At this setting the lens shows considerable distortions giving slightly bending wall grooves (see Figure 3). This was not calibrated or modelled. The camera moves in a distance of about 0.7 meter along a wall constructed from large roughly axed stones. The scene is roughly planar with deviations of about three centimetres. The camera was kept mostly normal to the surface – but free handed. This fairly well mimics the kind of videos that could be expected from an underwater vehicle cruising along a retaining wall. On the other hand, outside of the water we can easily step back and take a groundtruth picture with a longer focal length and less distortion. The one presented in Figure 6 was taken with a Pentax istD S using a standard SMC 1:2 f=35mm lens. Still this is not calibrated – however it is sufficiently free of distortions since this is not a zoom lens, and it can also be used

on the larger 35mm film frame. Moreover, only a section from the image centre is used.

Figure 3. One frame of the test video

Figure 4. Panorama patch from a hundred frames using standard homography estimation

Figure 5. Rectified panorama patch using homography decomposition following [3]

From the HD video taken with the Olympus PEN local panorama patches were stitched using the standard flow outlined in section 2.1 (Figure 1). Rather arbitrarily we set the number of frames to be composed into one patch to one hundred. One such patch can be seen in Figure 4. While the view seems fairly normal on the left hand side – where the initializing frame was – it is evident that the projective drift effects already start at the right hand side of the patch (the moment that no overlap is given). We can see the kind of problem homography stitching has by using our knowledge that the stones are truly rectangular – see groundtruth in the upper picture of Figure 6. A certain drift – in particular in the "projective" entries H_{31} and H_{32} is inevitable. Figure 3 shows the rectification of this patch using the decomposition method described in Section 2.2 on the homography corresponding to the patch. It finds a reasonable compromise correcting the mistakes. In particular the rectangular structure is reproduced better. Some shear drift remains. This rectified patch is than part of the larger panorama displayed as lower picture in Figure 6, which was obtained by the method indicated in Section 2.3. It can be seen that a beginning drift is sometimes corrected by force introducing considerable non-continuous steps into the homography chain.

4. CONCLUSION AND OUTLOOK

Here we could only present a very preliminary overview of the intended system. It was mainly tested on videos from outside water with mild distortions and rather good quality. Less favourable data can be expected from under water platforms. On such data often nothing can be seen. If something is seen the lighting may well be quite inhomogeneous, there will probably be floating clutter in front of the interesting scene, and the lens may be out of focus – autofocus should be off in order not to be disturbed by the clutter. Lately, we obtained such a video, and one frame of it is presented in Figure 7.

The processing chain as indicated in the activity diagrams above hast to be adjusted to such situations. The same parameter values, e.g. for the interest operator, as applied to the test sequence of section 3 (just the usual default settings) give less than four interest points sometimes and on many other occasions RANSAC still fails to come up with a plausible solution. The computational flow only took the "else" path of the partial homography chain diagram once in more than thousand images of the example video, while persistently staying on that side for hundreds of (successive) frames of the

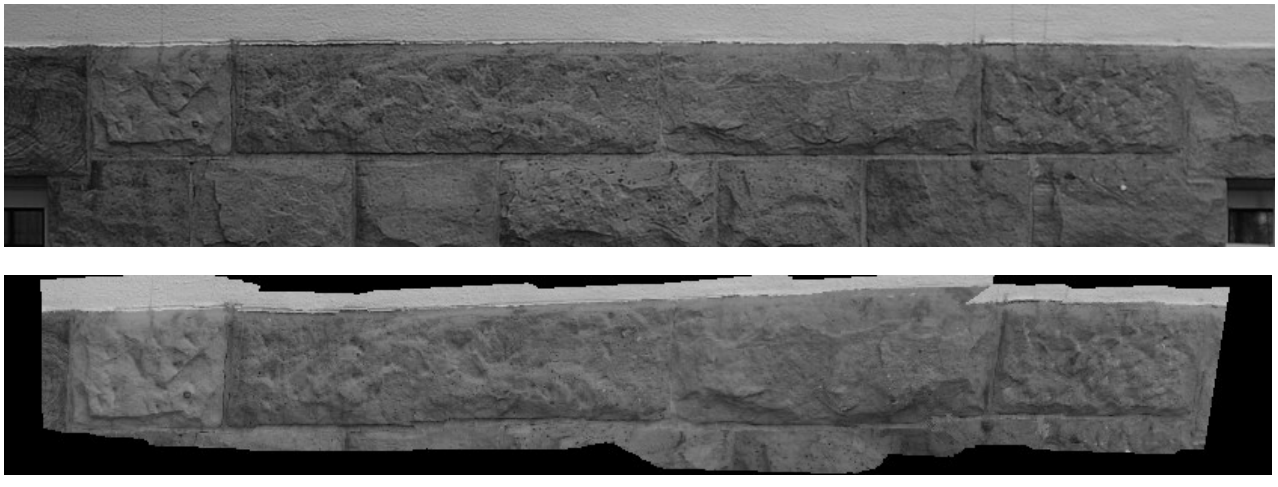

Figure 6: Lower: A panorama stitched from 5 patches, i.e. 500 frames, leading far away from the original frame on the leftmost side; Upper: Groundtruth picture of the same scene taken from further away

underwater video with the default parameters. Leaving the default settings in direction to more liberal ones of course gives less stable behaviour of the whole thing. Still, a preliminary result displayed in Figure 8 indicates sufficient stability to cope with such data. It is a good advice for underwater inspection to steer the vehicle as close to the structure of interest as possible.
It also becomes evident that the projective drift problem occurring when large sequences of such close-up videos are stitched can be mitigated by allowing full homography only on a local scale and keeping the global transform fixed to simple 2D-translation. The decomposition of the homography between the first and last frame of a patch giving an estimate of the surface orientation of the scene turns out to be an important help for rectification and subsequent joining of the patches.

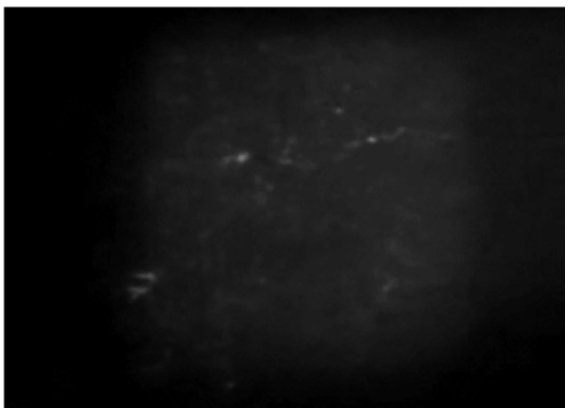

Figure 7: One frame of a typical underwater video

Obviously there is a trade-off between large patches from long camera movements allowing a stable decomposition with little error on the surface normal and epipole estimation on the one hand and the indicated projective drift problems that immediately begin to occur when there is no sufficient overlap anymore. Setting this parameter to a hundred frames can only be a first guess that has to be replaced by a mathematical investigation searching for the optimal patch size. Of course we look forward to making more experiments with challenging under water videos in the future. There remains a lot of room for improvement in all steps of the method.

References

[1] d'Angelo, P. HUGIN 2010.4.0, free panorama stitcher, http://hugin.sourceforge.net/download/ (accessed 28 Apr. 2011)

[2] Elibol, A., Moeller, B., Garcia, R., October 2008. Perspectives of Auto-Correcting Lens Distortions in Mosaic-Based Underwater Navigation. *Proc. of 23rd IEEE Int. Symposium on Computer and Information Sciences (ISCIS '08)*, Istanbul, Turkey, pp. 1-6.

[3] Faugeras, O., Lustman, F., 1988. Motion and structure from motion in a piecewise planar environment. *International Journal of Pattern Recognition and Artificial Intelligence*, 2(3), pp. 485–508.

[4] Fischler, M. A., Bolles, R. C., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the Association for Computing Machinery,* 24(6), pp. 381–395.

[5] Foerstner, W., 2010. Minimal Representations for Uncertainty and Estimation in Projective Spaces. *ACCV* (2), pp. 619-632

[6] Harris, C., Stephens, M., 1988. A combined corner and edge detector. *Proceedings of the 4th Alvey Vision Conference*. pp. 147–151. http://www.bmva.org/bmvc/1988/avc-88-023.pdf

[7] Hartley, R., Zisserman, A., 2000. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge.

[8] Köthe, U., 2003. Edge and Junction Detection with an Improved Structure Tensor. In: Michaelis, B., Krell, G. (Eds.): *Pattern Recognition, Proceedings* 25th*-DAGM*, Springer LNCS 2781, Berlin, pp. 25-32

[9] Laser Optronics, Underwater Gated Viewing Cameras, http://www.laseroptronix.se/gated/aqly.html, (accessed 28 Apr. 2011)

[10] Malis, E., Vargas M., Sep. 2007. Deeper understanding of the homography decomposition for vision-based control. INRIA report no. 6303, Sophia Antiopolis, France. http://hal.inria.fr/docs/00/17/47/39/PDF/RR-6303.pdf (accessed 28 Apr. 2011)

[11] Michaelsen, E., von Hansen, W., Kirchof, M., Meidow, J., Stilla, U., 2006. Estimating the Essential Matrix: GOODSAC versus RANSAC. In: Foerstner, W., Steffen, R. (eds) Proceedings Photogrammetric Computer Vision and Image Analysis. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Science,* Vol. XXXVI Part 3.

[12] Open CV Sources and DLLs, in particular http://opencv.willowgarage.com/documentation/cpp/video_motion_analysis_and_object_tracking.html?highlight=opticalflow#calcOpticalFlowPyrLK (accessed 26 Jun. 2011)

[13] Ren, Y., Chua, C.-S., Ho, Y.-H., 2003. Statistical background modeling for non-stationary camera. Pattern Recognition Letters (24), pp. 183–196

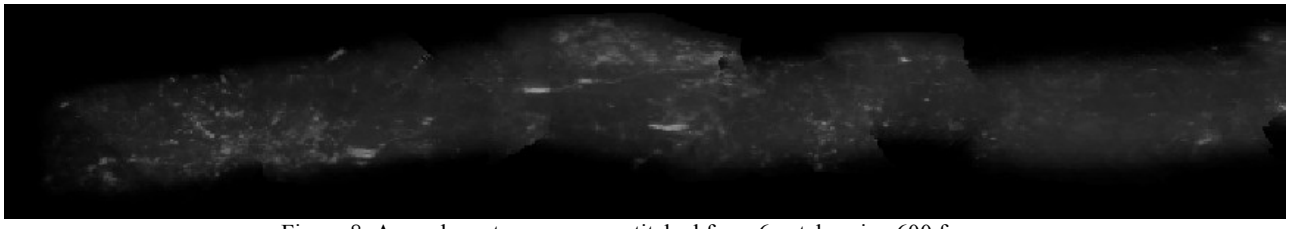

Figure 8: An underwater panorama stitched from 6 patches, i.e. 600 frames