

**Real-time Video Alignment and Fusion Using Feature Detection on FPGA
Devices**

A Thesis

Submitted to the Faculty

of

Drexel University

by

Robert Haywood Taglang

in partial fulfillment of the
requirements for the degree

of

Master of Science in Computer Engineering

June 2017



© Copyright 2017

Robert H. Taglang. All Rights Reserved.

Table of Contents

List of Tables	iii
List of Figures	iv
Abstract	v
1 Background	1
1.1 Introduction	1
1.2 Laplacian Fusion	2
1.3 Speeded-up Robust Features (SURF)	6
1.3.1 Computation of Hessian Determinants	6
1.3.2 SURF Implementations for FPGA Devices	9
1.4 Iterative Closest Point Algorithm	9
1.5 Singular Value Decomposition (SVD)	11
1.5.1 SVD Implementations for FPGA Devices	12
2 Implementation	13
2.1 Camera Interface	13
2.2 Streaming Kernel Operators	13
2.2.1 Hessian Kernel Operators	14
2.2.2 Average Filter Approximation of Single Level Laplacian Pyramid	16
2.2.3 Application of Non-maximum Suppression	16
2.3 Computation of Transform from Detected Features	16
2.4 Application of Transform to Real-Time Data	18
3 Results	18
3.1 Detecting feature points in real time	18
3.2 Selecting an optimal number of feature points	19
3.3 Implemented Design Details	20
4 Conclusions	20
References	27

List of Tables

1	Stages 1-8 of pipeline instructions for computing Hessian determinants	23
2	Stages 9-13 of pipeline instructions for computing Hessian determinants	24
3	Stages 14-16 of pipeline instructions for computing Hessian determinants	25
4	Timing data for the implemented design	25
5	Power data for the implemented design	26
6	Utilization data for the implemented design	26

List of Figures

1	Block diagram of the proposed system for alignment and fusion	1
2	Gaussian pyramid of an example image	3
3	Laplacian pyramid of an example image	4
4	Two images of the same scene with variations in sharpness and colorspace . .	5
5	Fusion of the images in Figure 4 using a naive selection and weighted sum approach	5
6	3D surface plots of the Gaussian second order derivative functions where $\sigma = 1$.	8
7	9×9 discrete approximations of the Gaussian second order derivative with $\sigma = 1.2$	9
8	Block diagram for computing 3×3 kernel convolution on a data stream using a scan chain	14
9	9×9 SURF kernels with marked integral sampling points	15
10	Bubble-sort architecture of the feature buffer	16
11	Feature buffer brute-force comparison	17
12	Feature buffer iterative comparison	18
13	Real-time capture of Hessian feature points	19
14	Images of "The Water Boy" by Frédéric-August Bartholdi from the Drexel Collection	20
15	The error in pixels per feature for performing n iterations of the iterative closest point algorithm on differently sized sets of the strongest feature points .	21
16	Alignment results after running iterative closest point for 16 iterations on varying numbers of features with reference features marked in red and transformed features in blue	22
17	The error in intensity based on the alignment from performing n iterations of the iterative closest point algorithm on differently sized sets of the strongest feature points	25
18	FPGA implemented design for live fusion using feature point detection . . .	26

Abstract

Real-time Video Alignment and Fusion Using Feature Detection on FPGA Devices

Robert Haywood Taglang

Prawat Nagvajara, Ph.D.

Video fusion functions as a way to combine the important or useful parts of two or more sequences of images. The scenario presented is the use of Laplacian fusion to produce a single video composed of the fields of view of two cameras whose areas of focus differ substantially. This is not a useful real-time strategy unless the fields can be aligned. This thesis presents a system for detecting features using an FPGA implementation of SURF (Speeded-Up Robust Features), and aligning video streams by applying a transform generated from the key features.

1 Background

1.1 Introduction

The fusion of data from two or more sensors has been well-researched [12] [10], though these approaches typically discuss the process of fusing images which have been pre-aligned. Pre-computed transforms used to align the frames of two cameras are not robust to variations. Some approaches have made use of additional hardware sensors in order to correct against these variations [5]. The approach presented in this thesis seeks to perform this correction in real time, completely in hardware, using feature detection on a Field Programmable Gate Array (FPGA) using the design shown in Figure 1.

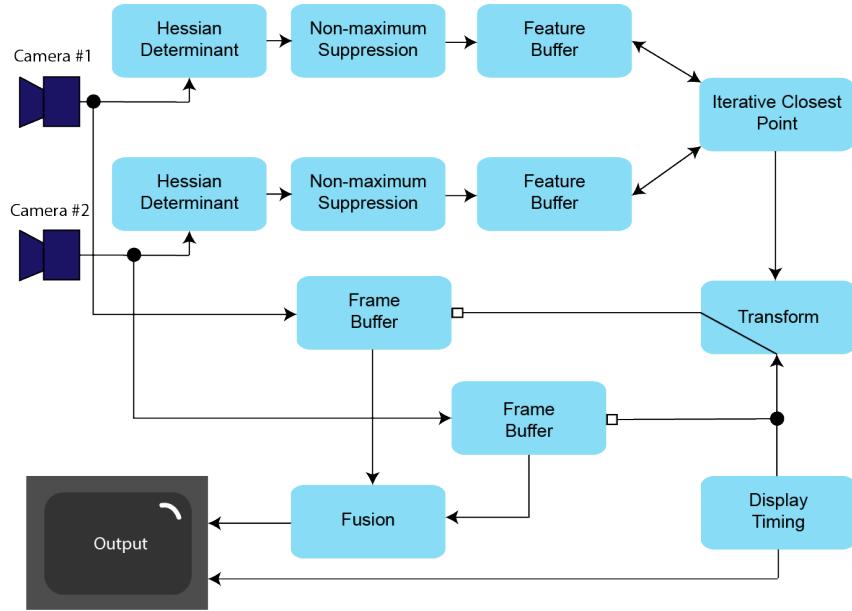


Figure 1: Block diagram of the proposed system for alignment and fusion

In order to align two images, one is treated as the reference image, and the other is the transform image which needs to be transformed to be aligned with the reference image. As such the problem can be broken into four major components.

1. Detecting feature points in the two images
2. Computing the transform which maps the transform image into the space of the reference image
3. Applying the computed transform to the transform image
4. Fusion of the reference and transformed image

Feature points are detected using some of the techniques from Speeded-up Robust Features (SURF) which makes use of Hessian determinants to detect points of interest in multiple scale spaces. High magnitude features are stored in a buffer to be used for computing the transform.

Once the feature points for the two images have been computed and stored in the buffer, the points from the transform image are mapped to their closest point in terms of euclidean distance from the reference set. An orthogonal projection from the transform points to the reference points is created with the transform as the unknown. Singular value decomposition is used to compute the pseudo-inverse of the matrix containing the reference points so it can be premultiplied by the matrix containing the transform points. The result of this product is the transform matrix which maps transform points to reference points with the error minimized in a least-squares sense. This process may need to be repeated multiple times to converge to a local error minimum, a process referred to as iterative closest point.

As the image data streams into hardware, the data is buffered into memory. With the transform computed, the desired address for the output image is decomposed into x and y coordinates, which are transformed and then reformed back into an address which is used to select pixels from the data in memory. In this way, the transform is applied to align the transform image with the reference image.

Finally, the technique of Laplacian fusion is used to combine the aligned images. It effectively selects the highest frequency components from the two images in order to create an output image where the sharpest, focused parts of the two images are combined into one image.

This process is applied continuously in real time. As the data from the camera data streams in, it produces a single fused output image for display with minimal latency. The design choice to use a FPGA rather than a GPU or some other software based approach was made due to the advantages gained from operating with custom embedded hardware over more rigid architectures. Overall, custom hardware implementations will always beat off-the-shelf components in terms of size, weight, and power consumption.

1.2 Laplacian Fusion

Laplacian pyramids of images have their origin as a strategy for image encoding [4]. A gaussian blur is applied to the image, and the image is downsampled to half of its original

size. In this context, Gaussian blur refers to applying kernel convolution to this image with a Gaussian kernel. Likewise, a box blur is the application of a uniformly valued, normalized kernel through kernel convolution. Downsampling refers the process of halving the resolution of the image by combining adjacent pixels. Upsampling is the process of doubling the resolution of the image and interpolating the missing values in the new image. The process of blurring and downsampling can be repeated on the resulting image to create a sequence of images representing the original in different scale spaces. This sequence of blurred and downsampled images is known as a Gaussian pyramid. An illustration of a Gaussian pyramid for an example image can be seen in Figure 2.

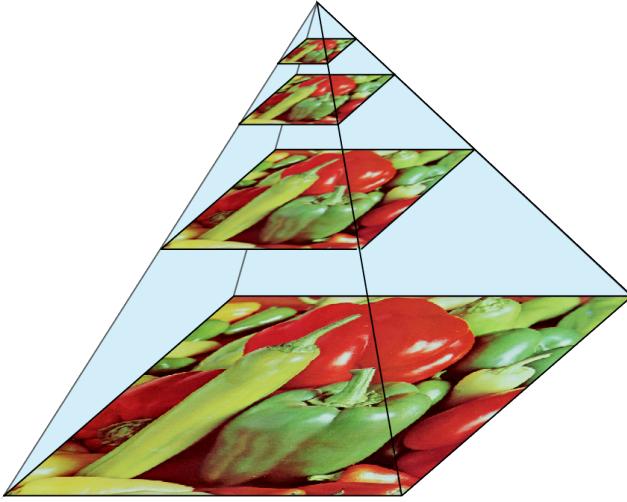


Figure 2: Gaussian pyramid of an example image

The Laplacian pyramid is one which can be used for reconstruction of the original image. At each level above the lowest level of the Gaussian pyramid, the level below is upsampled to match the scale of the current level. The difference between the upsampled image and the current scale level image is known as the Laplacian of the image. The sum of the upsampled lower level and the Laplacian is the original image. At a single level, the Laplacian can be thought of as the error introduced by applying a Gaussian and Box filter. A diagram illustrating a Laplacian pyramid can be seen in Figure 3. In Figure 3, the $G \downarrow$ operator refers to applying the Gaussian blur and downsampling, the \uparrow operator refers to upsampling the input image, and the $-$ operator refers to computing the difference of the two input images on each channel.

The property of the Laplacian that makes it ideal for fusion is its ability to capture the high frequency components in an image through the use of very simple kernel operators

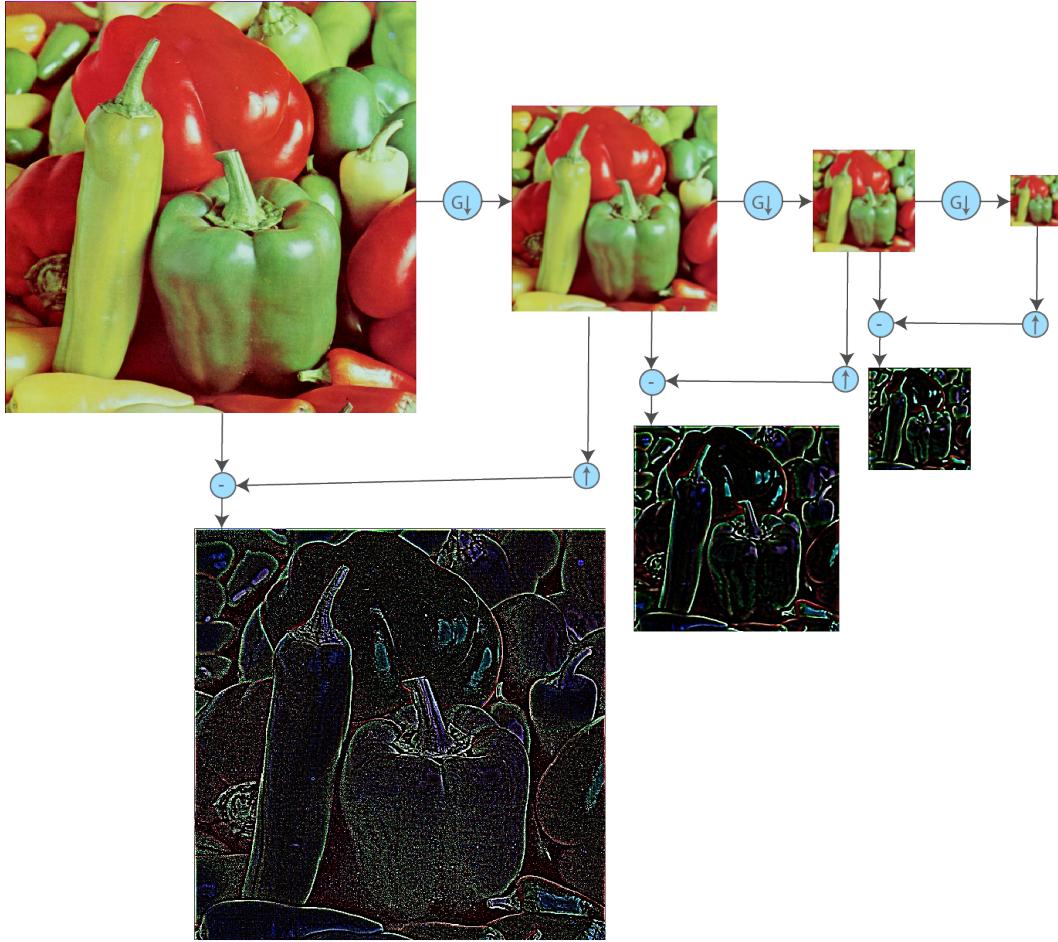


Figure 3: Laplacian pyramid of an example image

that are easily implemented in hardware. The difference between a blurred image and the original will have higher magnitude in the areas where the image was sharpest.

The fusion of two images can be thought of as a function of the two images X and Y of dimension $M \times N$ where $Z = f(X, Y)$, a single image of dimension $M \times N$. A naive approach to fusion would be to compute the Laplacians and use their magnitudes to select a pixel from either X or Y as shown in Equation 1. In this context, $|L(x)|$ represents the absolute value of the Laplacian of the input.

$$Z(i, j) = \begin{cases} X(i, j) & |L(X(i, j))| \geq |L(Y(i, j))| \\ Y(i, j) & otherwise \end{cases} \quad (1)$$

This approach does not account for variations in colorspace between the two images. Consider the images in Figure 4. The more saturated image will likely have a higher valued

Laplacian in some parts simply because it is brighter, therefore having higher magnitudes at individual pixels. This approach also will not facilitate smooth stitching of the images. Contiguous regions of selection from one image will be adjacent to regions from the other with no transition, producing a grainy effect at areas of high frequency. The result of this naive fusion can be seen in Figure 5a which exhibits the flaws of this approach.



(a) Blurred on the left; more saturated (b) Blurred on the right; more neutral

Figure 4: Two images of the same scene with variations in sharpness and colorspace



(a) Fusion using the naive approach (b) Fusion using weighted sum

Figure 5: Fusion of the images in Figure 4 using a naive selection and weighted sum approach

A more correct approach would involve using the Laplacian in a weighted sum to combine the pixels of the images, rather than simply selecting them, as shown in Equation 2.

$$Z(i, j) = \frac{|L(X(i, j))|}{|L(X(i, j))| + |L(Y(i, j))|} \cdot X(i, j) + \frac{|L(Y(i, j))|}{|L(X(i, j))| + |L(Y(i, j))|} \cdot Y(i, j) \quad (2)$$

The result of this weighted sum approach can be seen in Figure 5b which exhibits a reduction in graininess from the naive approach.

1.3 Speeded-up Robust Features (SURF)

The generation of features for use as marker points in alignment utilizes the SURF algorithm from Bay et al [2]. SURF is composed of two parts: a discrete approximation for computing Hessian determinants, and the generation of rotation invariant feature descriptors for detected feature points.

SURF is typically used for its applications in object recognition, where the feature descriptor is used to facilitate a match between what is observed and some known set of feature points and descriptors. The descriptor largely serves as a way of discriminating against false positives. In terms of using SURF for fusion, the detected feature points will be matched across two images with the assumption that the subject is the same and that the images do contain spatially coherent matches. Given this assumption, it can be concluded that the feature descriptor is not necessary for alignment. The feature points computed using Hessian determinants are sufficient.

1.3.1 Computation of Hessian Determinants

The Hessian determinant is the determinant of a Hessian matrix, which is a matrix composed of the spatial partial second derivatives of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. It is of the general form shown in Equation 3. In the image domain, $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. The particular form of the Hessian matrix in \mathbb{R}^2 with dimensions x_1 and x_2 is shown in Equation 4.

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \quad (3)$$

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix} \quad (4)$$

The second order derivative used can be computed via convolution of the Gaussian second order derivative at any point, x , in the image. The formulas for the Gaussian second order derivative for each partial with respect to x_1^2 , x_1x_2 and x_2^2 can be seen in Equations 5, 6, and 7 respectively.

$$\frac{\partial^2 G(x_1, x_2, \sigma)}{\partial x_1^2} = (-1 + \frac{x_1^2}{\sigma^2}) \frac{e^{-\frac{x_1^2+x_2^2}{2\sigma^2}}}{2\pi\sigma^4} \quad (5)$$

$$\frac{\partial^2 G(x_1, x_2, \sigma)}{\partial x_1 \partial x_2} = \frac{x_1 x_2}{2\pi\sigma^6} e^{-\frac{x_1^2+x_2^2}{2\sigma^2}} \quad (6)$$

$$\frac{\partial^2 G(x_1, x_2, \sigma)}{\partial x_2^2} = (-1 + \frac{x_2^2}{\sigma^2}) \frac{e^{-\frac{x_1^2+x_2^2}{2\sigma^2}}}{2\pi\sigma^4} \quad (7)$$

3D surface plots of these equations where $\sigma = 1$ can be seen in Figure 6. In order to compute these functions quickly, SURF approximates them with cropped, discrete, kernels.

The Gaussian second order derivatives can be approximated as 9×9 kernels with $\sigma = 1.2$. These kernels can be seen in Figure 7. By adjusting σ , Hessian determinants can be computed in different scale spaces. The approximation of H with 9×9 kernels is apparent when comparing Figure 6 and Figure 7. This is a concept that SURF draws from the Scale-Invariant Feature Transform (SIFT) from Lowe et al [11].

By detecting features in different scale spaces, SIFT and by proxy, SURF are robust to changes in scale. SIFT accomplished this by downsampling the image to detect at lower order scale spaces. SURF improved on this approach for speed by instead scaling the kernel.

Another speed optimization presented in SURF takes advantage of the form of the discrete kernels. Since the approximated kernels are composed of rectangles of constant value, they can be decomposed into a set of box filters. Box filters can be computed quickly with the use of integral images. The general form of the integral image of an $M \times N$ image I is shown in Equation 8.

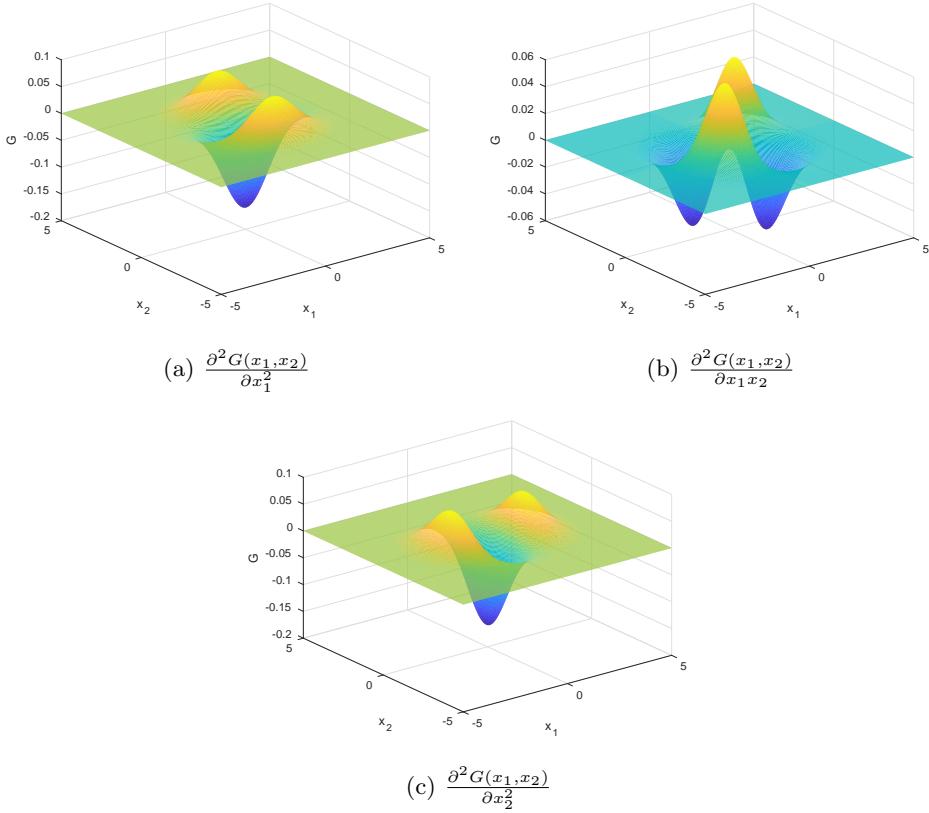


Figure 6: 3D surface plots of the Gaussian second order derivative functions where $\sigma = 1$

$$\int I = \begin{bmatrix} I(0, 0) & \sum_{m=0}^1 I(0, m) & \dots & \sum_{m=0}^M I(0, m) \\ \sum_{n=0}^1 I(n, 0) & \sum_{n=0}^1 \sum_{m=0}^1 I(n, m) & & \vdots \\ \vdots & & \ddots & \vdots \\ \sum_{n=0}^N I(n, 0) & \dots & \dots & \sum_{n=0}^N \sum_{m=0}^M I(n, m) \end{bmatrix} \quad (8)$$

Integral images decrease the computational complexity of finding the sum of an area in the input image. The computational complexity of strict kernel convolution at a point scales with the size of the kernel and is of the order $O(N^2)$ where the kernel is $N \times N$. A box filter can be decomposed into finding the sum of an area in the image and scaling it. Finding the sum from the integral image can be performed in $O(1)$ as shown in Equation 9.

$$\sum_{n=w}^y \sum_{m=x}^z = \int I(y, z) - \int I(w-1, z) - \int I(y, x-1) + \int I(w-1, x-1) \quad (9)$$

This makes the computation of the Hessian matrix scale only in terms of the image size, yielding no additional penalty for operating on different scale spaces.

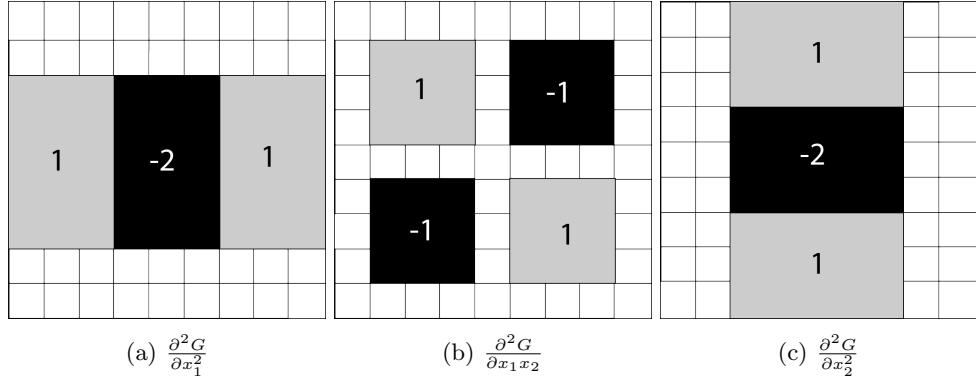


Figure 7: 9×9 discrete approximations of the Gaussian second order derivative with $\sigma = 1.2$

1.3.2 SURF Implementations for FPGA Devices

The relatively low computational complexity makes SURF a popular choice for FPGA applications. Battezzati et al. present an architecture using accumulators for computing the integral image pipelined through the Hessian computation and storing detected feature points in a first in, first out (FIFO) cache [1]. These are matched against a stored set of feature points. Chen et al. present improvements on this approach by parallelizing the computation of different scale spaces [6]. The implementation in this thesis follows these approaches with some additional improvements for speed based on the use case of matching against another image rather than a stored set of features.

1.4 Iterative Closest Point Algorithm

The crucial step in aligning the two images is the computation of an affine transform mapping one image into the space of the other. Once the images are aligned, they can be fused. A combination of SURF and the iterative closest point algorithm are used to compute this transform. Iterative closest point was first proposed by Chen and Medioni as a method for aligning 3-D point cloud data [7]. In its simplest form, the algorithm follows the following steps:

1. Each point in the set of points to be transformed is matched against the closest (usually Euclidean distance) point in the reference set of points.
2. A transformation is estimated to minimize the distance between the transform set and their matches in the reference set of points.
3. The transform is applied to the points.

This process is repeated, converging to the local minimum that is the match between the two point sets.

Let $X_p(i)$ be the i^{th} point in the set of reference points onto which the transform points, $X(j)$, where j is the index of the closest point to $X_p(i)$ in X , will be projected. The general form for this transformation M is shown in Equation 10, and the expanded matrix form can be seen in Equation 11. In the expanded matrix form M is decomposed into R , a 2×2 rotation matrix, and T , a translation offset.

$$X_p(i) = X(j) \cdot M \quad (10)$$

$$\begin{bmatrix} X_p(i)_1 \\ X_p(i)_2 \\ 1 \end{bmatrix} = \begin{bmatrix} X(j)_1 & X(j)_2 & 1 \end{bmatrix} \cdot \begin{bmatrix} R_{11} & R_{12} & T_1 \\ R_{21} & R_{22} & T_2 \\ 0 & 0 & 1 \end{bmatrix} \quad (11)$$

This equation only solves for M for a single point relation, but can be restructured to contain the whole set of N points as shown in Equation 12. In this equation P is a function $P : i \rightarrow j$ mapping the closest points in each set.

$$\begin{bmatrix} X_p(0)_1 \\ X_p(0)_2 \\ X_p(1)_1 \\ X_p(1)_2 \\ \vdots \\ X_p(N)_1 \\ X_p(N)_2 \end{bmatrix} = \begin{bmatrix} X(P(0))_1 & X(P(0))_2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & X(P(0))_1 & X(P(0))_2 & 1 \\ X(P(1))_1 & X(P(1))_2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & X(P(1))_1 & X(P(1))_2 & 1 \\ \vdots & \vdots & & & \vdots & \\ X(P(N))_1 & X(P(N))_2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & X(P(N))_1 & X(P(N))_2 & 1 \end{bmatrix} \cdot \begin{bmatrix} R_{11} \\ R_{12} \\ T_1 \\ R_{21} \\ R_{22} \\ T_2 \end{bmatrix} \quad (12)$$

Solving for M in this way maps all points in X to their closest points in X_p based on an orthogonal projection with the error minimized in a least-squares sense.

In this form, the transform M will include shear transformations and non-uniform scaling. The computation can be simplified by forcing the second basis vector in R to be orthogonal to the first. The two cameras are expected to be physically in the same plane, and as such, non-uniform scaling and shear transformations are not expected for alignment. By setting $R_{21} = -R_{12}$ and $R_{22} = R_{11}$, the computation of M can be reduced

as shown in Equation 13.

$$\begin{bmatrix} X_p(0)_1 \\ X_p(0)_2 \\ X_p(1)_1 \\ X_p(1)_2 \\ \vdots \\ X_p(N)_1 \\ X_p(N)_2 \end{bmatrix} = \begin{bmatrix} X(P(0))_1 & X(P(0))_2 & 1 & 0 \\ X(P(0))_2 & -X(P(0))_1 & 0 & 1 \\ X(P(1))_1 & X(P(1))_2 & 1 & 0 \\ X(P(1))_2 & -X(P(1))_1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ X(P(N))_1 & X(P(N))_2 & 1 & 0 \\ X(P(N))_2 & -X(P(N))_1 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} R_{11} \\ R_{12} \\ T_1 \\ T_2 \end{bmatrix} \quad (13)$$

If scaling is disallowed, and the transformation consists of only a translation and a rotation, this can be reduced further. Consider first computing the centroids of the point sets as in Equation 14.

$$C = \frac{1}{N} \sum_{i=1}^N X(i) \quad (14)$$

Based on the closest point matching, the covariance matrix H can be computed as in Equation 15.

$$H = \sum_{i=1}^N (X(P(i)) - C) \cdot (X_p(i) - C_p)^T \quad (15)$$

The singular value decomposition $U\Sigma V = SVD(H)$ can be used to compute the rotation $R = VU^T$, where the translation is the distance between the centroids.

The performance of iterative closest point can be further improved by making it more sensitive to errors. Chetverikov introduced a variant of iterative closest point referred to as trimmed iterative closest point (TrICP) [8]. TrICP is more robust to errors by eliminating points that introduce error into the matching. Some detected features points will not have correspondences between images. By eliminating these points, the overall error can be reduced to compute a more accurate transform.

1.5 Singular Value Decomposition (SVD)

Let Equation 13 be of the form $X = Q \cdot M$. In order to compute the transform M , the equation must be restructured as in Equation 16.

$$Q^{-1}X = M \quad (16)$$

Q is not a square matrix, and as such is not invertible, but its pseudoinverse can be used in this instance. The pseudoinverse is computed through the use of singular value decomposition. The use of singular value decomposition to compute the transform is the source of the least squares fitting achieved by iterative closest point.

This is necessary for computing transforms with more degrees of freedom, however for rigid body transformations, it is sufficient to find the singular value decomposition of the covariance matrix H , a 2×2 matrix for which a closed form solution does exist[3] and is shown in Equation 17.

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix} &= \begin{bmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{bmatrix} \begin{bmatrix} w_1 & 0 \\ 0 & w_2 \end{bmatrix} \begin{bmatrix} \cos\gamma & \sin\gamma \\ -\sin\gamma & \cos\gamma \end{bmatrix} = \begin{bmatrix} E & H \\ -H & E \end{bmatrix} + \begin{bmatrix} F & G \\ G & -F \end{bmatrix} \\ \frac{w_1 + w_2}{2} &= \sqrt{E^2 + H^2} \\ \frac{w_1 - w_2}{2} &= \sqrt{F^2 + G^2} \\ \gamma - \beta &= \tan^{-1}(G/F) \\ \gamma + \beta &= \tan^{-1}(H/E) \end{aligned} \quad (17)$$

1.5.1 SVD Implementations for FPGA Devices

Singular value decomposition can be performed on FPGAs by cascading a set of 2×2 cells [13]. Ledesma-Carillo et al. present a hardware efficient algorithm for computing singular value decompositions on large matrices using one-sided Jacobi rotations for computing SVD on arbitrary $M \times N$ matrices [9]. One of these approaches is necessary if the sensor fusion must correct for scale or shear. It is worth noting that these approaches both require very high utilization of the FPGA, and as such may be difficult to implement for large numbers of feature points.

In contrast, if only a rigid body transformation is required, then Equation 17 can be implemented trivially using CORDIC approximations for the square root and arctangent functions.

2 Implementation

The implementation of this design was performed on a ZedBoard Zynq-7000 ARM/FPGA SoC Development Board.

2.1 Camera Interface

The design was implemented using a pair of OV7670 VGA cameras. These cameras feature an I^2C interface for configuration, and generate hsync and vsync VGA timing signals along with 8 bits of data. They can be configured to output 16-bit RGB(565) with half of the RGB signal sent on each clock pulse. Configured this way, the camera outputs a resolution of 640 by 480 pixels at 30Hz.

2.2 Streaming Kernel Operators

Performing kernel convolution in real time is complicated by the fact that a pixel has data dependencies on its neighbors. As such, the input data must be buffered until all of the necessary data is available to perform convolution at a given point. The buffer size must be at least $N \times W + M$ where the kernel is $M \times N$ and the image is of width W . On FPGAs, this minimally sized buffer can be implemented using LUTRAM, a hardware lookup table. LUTRAM on most FPGAs also has the advantage of being able to act as a set of shift registers. The convolution multipliers and adders can be attached to a single set of cells, and the data can be shifted through the array as it streams. The architecture for this scan chain approach for computing 3×3 kernel convolution can be seen in Figure 8.

This scan chain approach to kernel convolution is the faster and more utilization efficient approach to performing kernel convolution on streamed data. However, LUTRAM on most FPGAs is a limited resource, and as such, large kernels and image widths will make it difficult to implement a design. Implementing an LUTRAM scan chain for performing operations on 9×9 kernels with an image width of 640 pixels on the Xilinx Zynq XC7Z020-1CLG484 chipset used for this design consumed 96% of the available LUTRAM. Since it was necessary to implement two hessian operators, one for each camera, an approach was devised to utilize the slower, but more plentiful block RAM.

The block RAM has limitations. The data cells inside of the block RAM are not directly accessible. They must be read or written by writing an address to a control port and then waiting for the designated latency. The process of reading and writing to and from block

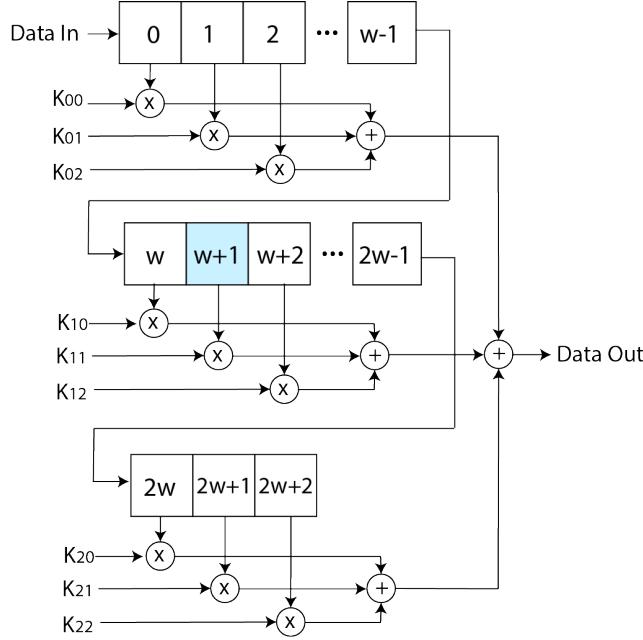


Figure 8: Block diagram for computing 3×3 kernel convolution on a data stream using a scan chain

RAM can be pipelined, that is, instructions can be given in sequence and then after the initial latency, data will be read and written every clock cycle. This will be explored in more detail in the design of the Hessian kernel operator.

2.2.1 Hessian Kernel Operators

The kernels discussed in this section are the 9×9 discrete approximations shown in Figure 7. If straightforward kernel convolution were to be implemented, it would require sampling all points with non-zero kernel values which would not be realistic for performing these operators in real time. Instead, the integral image is used to reduce the number of required sampling points. Recall that the integral image can be used to compute the area in a rectangle by sampling the corners, adding the bottom-right corner and top-left corner, and subtracting the bottom-left corner and top-right corner. As such, the required sampling points for the SURF kernels are highlighted in Figure 9.

It can be observed in Figure 9 that the worst-case for the data dependencies is the bottom-right point in $\frac{\partial^2 G}{\partial x_2^2}$. If the Hessian determinant is to be computed with minimal latency, it should be computed when that point becomes available. In order to compute the integral at a point, the integral above, to the left, and to the top-left of the desired point must be sampled. From this, it can be concluded that thirty-four points must be sampled

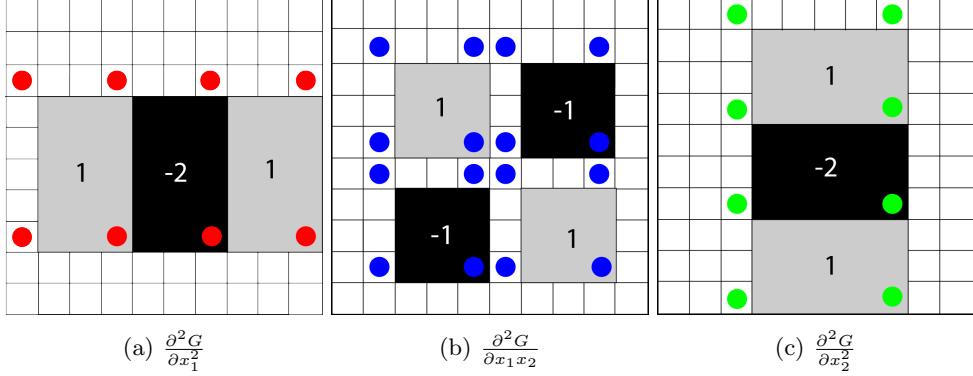


Figure 9: 9×9 SURF kernels with marked integral sampling points

to compute the Hessian determinant (eight from $\frac{\partial^2 G}{\partial x_1^2}$, sixteen from $\frac{\partial^2 G}{\partial x_1 x_2}$, seven from $\frac{\partial^2 G}{\partial x_2^2}$, and three to compute the integral).

In order to do this in real time, the Hessian determinant must be computable within a pixel clock cycle. Given a resolution of 640 by 480 at 30 frames per second, this is a pixel clock speed of $12.5MHz$. Recall that block RAM must be addressed and read. In this implementation, block RAM has a latency of three clock cycles. In order to compute the determinant in a single pixel clock cycle, the Hessian block is clocked at $200MHz$, giving it 16 clock cycles for every pixel clock cycle. A $200MHz$ clock has a period of only $5ns$ which makes pipelining of instructions very important in order to meet timing. A dual port block RAM was used, with addressing using the lower bits of y to map into a modular address space of 16 rows. The dual port RAM effectively allows for 32 read/write operations within the pixel clock period. An LUTRAM cache containing the integral values for the last row up to the top-left point required for the integral image reduces the number of memory reads for required points to 29, along with a single memory write to place the last value in the cache into the block RAM. The set of pipeline instructions for computing the Hessian determinants can be found in Tables 1, 2, and 3. In these tables, L represents the rectangles that make up the SURF kernels. Its subscripts, such as L_{xx_n} refer to the specific rectangle. L_{xx} is numbered from right to left, L_{xy} is numbered clockwise beginning at the top left, and L_{yy} is numbered from top to bottom referring to what is shown in Figure 9.

When computed in this way, the Hessian determinant at a point will be outputted at the end of the next pixel clock cycle continuously as the data stream passes through the system. It is worth noting that this implementation does not need to be modified to compute Hessian determinants in different scale spaces. Larger kernels still have the same number of points

that must be sampled, but the required block RAM for the design does grow with the kernel size.

2.2.2 Average Filter Approximation of Single Level Laplacian Pyramid

In this design, only a single level of the Laplacian pyramid is used for fusion. In a single level, a Gaussian blur is applied to the image. It is downsampled, upsampled, and subtracted from the original. The downsampling and upsampling in a single level can be approximated as an average filter without performing the costly operation of modifying the resolution of the image. In this design, the Gaussian and average filters were implemented as a set of scan chains to perform 3×3 kernel convolution.

2.2.3 Application of Non-maximum Suppression

Though not a kernel operator, non-maximum suppression depends on the data points around it. Non-maximum suppression is to be applied to the Hessian determinant values in order to reduce them to a smaller, more precise set of features. In a 3 by 3 neighborhood, if the central pixel is not the maximum, it is set to zero. In this way, only the peak features remain for processing. The 3×3 neighborhood search was implemented as a 3×3 scan chain with comparison operators.

2.3 Computation of Transform from Detected Features

As features are generated, they are passed into a feature buffer as a tuple of their magnitude, x , and y coordinates. The buffer is made up of bubble-sort cells as shown in Figure 10.

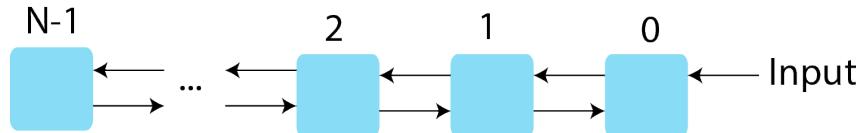


Figure 10: Bubble-sort architecture of the feature buffer

On the rising edge of the clock, the even cells swap values with their neighbor to the left if the cell contains a greater value than its neighbor. On the falling edge, the odd cells do the same. In this way, the highest valued elements propagate to the top of the buffer, and lower valued elements are dropped out of the bottom. The comparison criteria for this operation is the Hessian magnitude, and the sort is active for as long as features are being

generated. At the end of the frame, the buffer contains only the points with the highest Hessian magnitudes.

Once the points have been collected, the transform set is compared to the reference set and correspondences are assigned based on euclidean distance to find the closest point. This is done via a brute-force search. Every point is compared to every other, and the closest one emerges. The architecture of this comparison is shown in Figure 11.

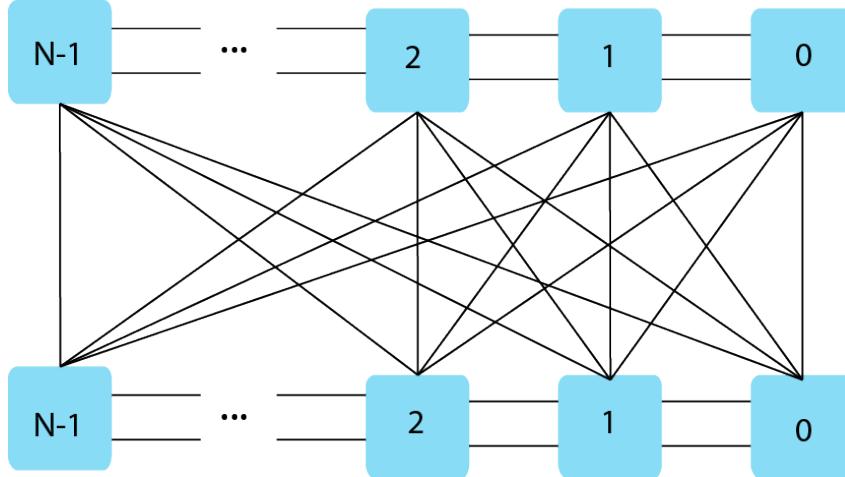


Figure 11: Feature buffer brute-force comparison

For low numbers of feature points, the inefficiency of brute-force search is mitigated by its simplicity. In hardware, a brute force comparison like this does not suffer from additional time complexity since the comparisons all happen in parallel. However, the space complexity and fanout of the circuit goes up exponentially as more cells are added.

It should be intuitive that this architecture is only useful for small numbers of feature points. The high fanout of the brute force comparison quickly becomes unmanageable to route in most designs with more than a handful of features. In the case that this becomes difficult, it is possible to trade the space complexity of the design for time complexity by making the computation iterative as in Figure 12.

In this approach, the a point from the reference set is loaded into register A. Then, each point in the transform set is loaded into register B in turn. If the distance between the loaded value and the value in register A is less than the distance between the value stored in the match register and register A, then the loaded value from register B is placed into the match register. This is repeated for all elements in the reference set to find their closest points in the transform set.

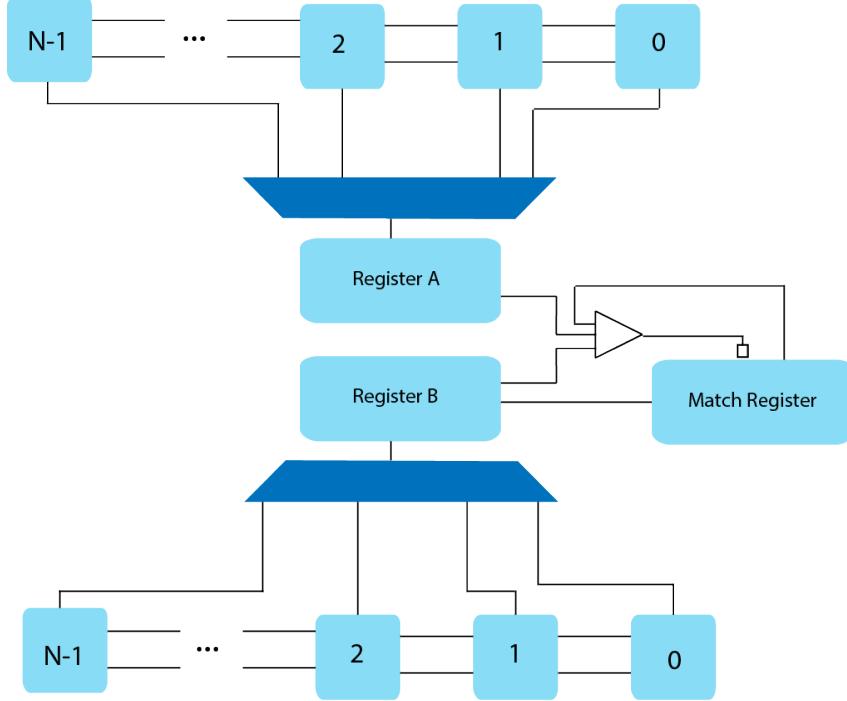


Figure 12: Feature buffer iterative comparison

2.4 Application of Transform to Real-Time Data

The transform is applied through the use of a frame buffer. The reference image remains fixed and is outputted as normal. However, the address for the transform image is decomposed into x and y coordinates. This is multiplied through the computed transform, and the transformed coordinates are used to select pixels from the frame buffer. The reference pixel and transform pixel are then fused together to create the output stream.

The transform is recomputed at the end of the frame from the detected features and then applied to the next generation of incoming data. In this way, the transform is additively refined as in iterative closest point. The transform computed for a frame is concatenated with the transform from the last frame, eventually converging to a local error minimum for alignment in a least squares sense.

3 Results

3.1 Detecting feature points in real time

The system for detecting feature points in real time is demonstrated through a picture of the system output shown in Figure 13. Detected feature points above a threshold are marked

in white, showing the detection of strong changes in gradient, particularly evident around the eyes and nose.



Figure 13: Real-time capture of Hessian feature points

3.2 Selecting an optimal number of feature points

A set of experiments to align the images shown in Figure 14 were performed using different numbers of feature points for a variable number of iterations of the iterative closest point algorithm.

The iterative closest point algorithm always converges to a solution which minimizes the overall distance between the transform and reference set of points. This can be observed in Figure 15 which shows the average distance from each point in the transform set of points to its closest neighbor in the reference set of points. Regardless of the selected number of features, all experiments converge to a roughly the same minimum. However, this may be an optimal solution in terms of distances between the feature points, but it is not always optimal in terms of alignment.

Figure 16 shows the alignment between the transform and reference image when certain numbers of feature points are selected and used to compute the transform. It can be observed that 20 features is not sufficient to compute a mapping, but that all of the images with more than 40 features are imperfectly aligned. This is in spite of the fact that iterative closest point converges to a local minimum in terms of displacement error.

From this, it can be concluded that selecting a small, but sufficient number of feature points produces optimal alignment. This is likely due to the fact that additional feature

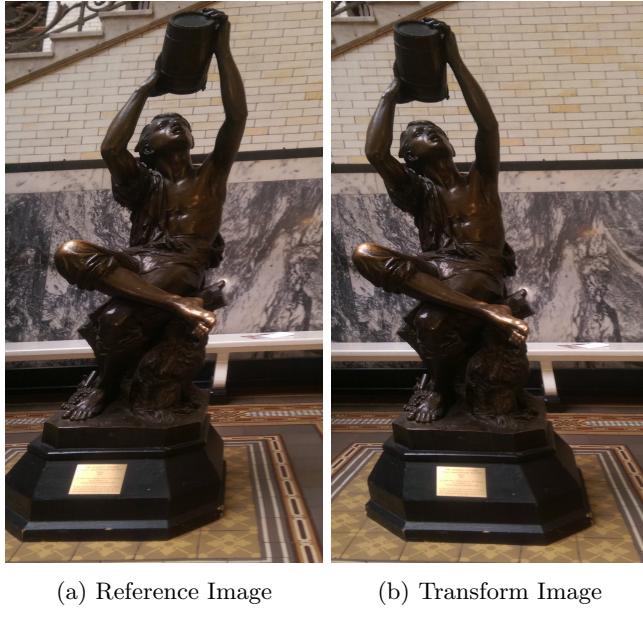


Figure 14: Images of "The Water Boy" by Frédéric-Auguste Bartholdi from the Drexel Collection

points beyond just the strongest features in the images introduce error by trying to align the noisier parts of the background. A better error metric may be the one shown in Figure 17 which shows the error in terms of the intensity difference. The magnitude of the overlaid image subtracted from the original is used to compute this error metric which more accurately reflects the alignment results shown in Figure 16.

3.3 Implemented Design Details

The implemented design for detecting feature points and transform the data for alignment and fusion can be seen in Figure 18. The timing, power, and utilization data can be found in Tables 4, 5, and 6 respectively.

4 Conclusions

The approach presented is capable of performing the task stated in the abstract in real time. Careful pipelining in a novel architecture for the computing of the hessian determinants makes this possible. The treatment of each frame as a generation of feature points to be transformed to a local error minimum as opposed to a more rigid architecture of performing multiple iterations on one feature set helps alleviate the overhead of the design, and make it realizable. The available chipset was not of sufficient size to perform the larger corrections

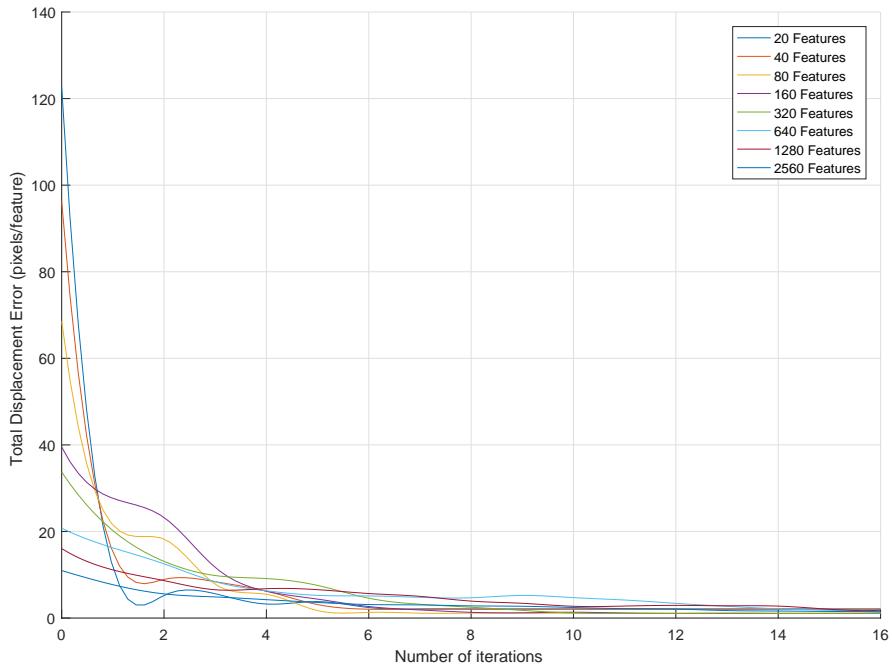


Figure 15: The error in pixels per feature for performing n iterations of the iterative closest point algorithm on differently sized sets of the strongest feature points

of scale and shear due to the prohibitive nature of instantiating large arrays of SVD solving cells. However, on a more robust chipset, scale and shear should also be correctable factors, as well as giving the ability to compute the transform from a larger set of feature points.



Figure 16: Alignment results after running iterative closest point for 16 iterations on varying numbers of features with reference features marked in red and transformed features in blue

Table 1: Stages 1-8 of pipeline instructions for computing Hessian determinants

Cycle	Instruction #1	Instruction #2
1.	<ul style="list-style-type: none"> • Store current (x,y) and value in registers 	<ul style="list-style-type: none"> • Get integral top, left, and corner from cache • Compute top left and top right addresses for L_{yy_2}/L_{yy_1} • Write addresses to read top left and top right of L_{yy_1}/L_{yy_0} • Data ready, bottom right and top right corner of L_{xx_2}/L_{xx_1} • Compute L_{xx_0}
2.	<ul style="list-style-type: none"> • Compute (x,y) bottom right and top left corner for L_{xy_0} 	<ul style="list-style-type: none"> • Write addresses to read top left and top right corner of L_{yy_2}/L_{yy_1} • Data ready, bottom left and top left corner of L_{xx_2} • Compute L_{xx_1} • Compute integral
3.	<ul style="list-style-type: none"> • Compute (x,y) bottom left and top right corner for L_{xy_0} Compute bottom right and top left addresses for L_{xy_0} 	<ul style="list-style-type: none"> • Data ready, top left and top right corner for L_{yy_0} • Compute L_{xx_2}
4.	<ul style="list-style-type: none"> • Compute (x,y) bottom right and top left corner for L_{xy_1} • Compute bottom left and top right addresses for L_{xy_0} • Write addresses to read bottom right and top left of L_{xy_0} 	<ul style="list-style-type: none"> • Data ready, top left and top right corner of L_{yy_1}/L_{yy_0} • Compute L_{xx}
5.	<ul style="list-style-type: none"> • Compute (x,y) bottom left and top right corner for L_{xy_1} • Compute bottom right and top left addresses for L_{xy_1} • Write addresses to read bottom left and top right of L_{xy_0} 	<ul style="list-style-type: none"> • Data ready, top left and top right corner of L_{yy_2}/L_{yy_1} • Compute L_{yy_0}
6.	<ul style="list-style-type: none"> • Compute (x,y) bottom right and top left corner for L_{xy_2} • Compute bottom left and top right addresses for L_{xy_1} • Write addresses to read bottom right and top left of L_{xy_1} 	<ul style="list-style-type: none"> • Compute L_{yy_1}
7.	<ul style="list-style-type: none"> • Compute (x,y) bottom left and top right corner for L_{xy_2} • Compute above point address for write • Compute top left address for L_{xy_2} • Write addresses to read bottom left and top right of L_{xy_1} • Data ready, bottom right and top left corner of L_{xy_0} 	<ul style="list-style-type: none"> • Compute L_{yy_2}
8.	<ul style="list-style-type: none"> • Compute (x,y) bottom right and top left corner for L_{xy_3} • Compute bottom left and top right addresses for L_{xy_2} • Write addresses to write above point and read top left corner of L_{xy_2} • Data ready, bottom left and top right corner of L_{xy_0} 	<ul style="list-style-type: none"> • Shift cache • Compute L_{yy}

Table 2: Stages 9-13 of pipeline instructions for computing Hessian determinants

9.	<ul style="list-style-type: none"> • Compute (x,y) bottom left and top right corner for L_{xy_3} • Compute bottom right and top left addresses for L_{xy_3} • Write addresses to read bottom left and right of L_{xy_2} • Data ready, bottom right and top left corner of L_{xy_1} • Compute L_{xy_0} 	
10.	<ul style="list-style-type: none"> • Compute (x,y) bottom right and top right for L_{xx_0} • Compute bottom left and top right addresses for L_{xy_3} • Write addresses to read bottom right and top left of L_{xy_3} • Data ready, bottom left and top right corner of L_{xy_1} 	<ul style="list-style-type: none"> • Compute determinant secondary diagonal
11.	<ul style="list-style-type: none"> • Compute (x,y) bottom right and top right for L_{xx_1}/L_{xx_0} • Compute bottom right and top right addresses for L_{xx_0} • Write addresses to read bottom left and top right of L_{xy_3} • Data ready, bottom right and top left corner of L_{xy_2} • Compute L_{xy_1} 	
12.	<ul style="list-style-type: none"> • Compute (x,y) bottom right and top left for L_{xx_2}/L_{xx_1} • Compute bottom right and top right addresses for L_{xx_1}/L_{xx_0} • Write addresses to read bottom right and top right for L_{xx_0} • Data ready, bottom left and top right corner of L_{xy_2} 	
13.	<ul style="list-style-type: none"> • Compute (x,y) bottom left and top left for L_{xx_2} • Compute bottom right and top right addresses for L_{xx_2}/L_{xx_1} • Write addresses to read bottom right and top right of L_{xx_1}/L_{xx_0} • Data ready, bottom right and top left corner of L_{xy_3} • Compute L_{xy_2} 	<ul style="list-style-type: none"> • Compute determinant primary diagonal

Table 3: Stages 14-16 of pipeline instructions for computing Hessian determinants

14.	<ul style="list-style-type: none"> • Compute (x,y) top left and top right for L_{yy_0} • Compute bottom left and top left addresses for L_{xx_2} • Write addresses to read bottom right and top right of L_{xx_2}/L_{xx_1} • Data ready, bottom left and top right corner of L_{xy_3} 	<ul style="list-style-type: none"> • Compute determinant
15.	<ul style="list-style-type: none"> • Compute (x,y) top left and top right for L_{yy_1}/L_{yy_0} • Compute top left and top right addresses for L_{yy_0} • Write addresses to read bottom left and top left of L_{xx_2} • Data ready, bottom right and top right corner of L_{xx_0} • Compute L_{xy_3} 	<ul style="list-style-type: none"> • Absolute value determinant
16.	<ul style="list-style-type: none"> • Compute (x,y) top left and top right for L_{yy_2}/L_{yy_1} • Compute top left and top right addresses for L_{yy_1}/L_{yy_0} • Write addresses to read top left and top right corner of L_{yy_0} • Compute L_{xy} 	<ul style="list-style-type: none"> • Write determinant

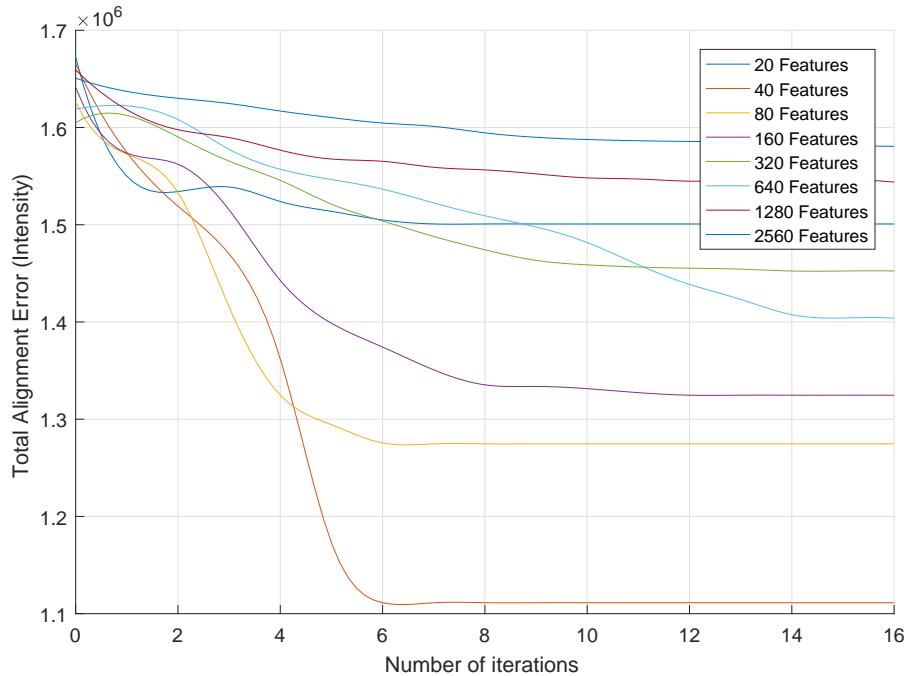


Figure 17: The error in intensity based on the alignment from performing n iterations of the iterative closest point algorithm on differently sized sets of the strongest feature points

Table 4: Timing data for the implemented design

Worst Case Slack (Setup)	$+0.081ns$
Worst Case Slack (Hold)	$+0.047ns$
Total # of Endpoints	4495

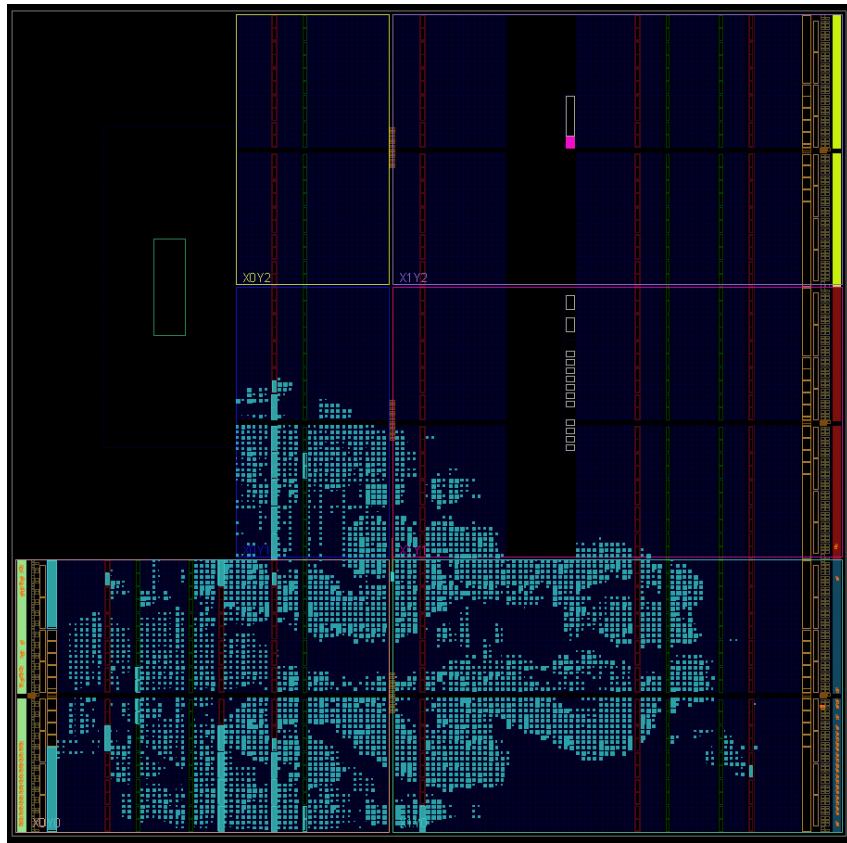


Figure 18: FPGA implemented design for live fusion using feature point detection

Table 5: Power data for the implemented design

Total On-Chip Power	$0.615W$
Junction Temperature	$32.1^{\circ}C$

Table 6: Utilization data for the implemented design

Look-up Table	8758
Flip-flops	6645
Block-RAM	22

References

- [1] N. Battezzati, S. Colazzo, M. Maffione, and L. Senepa. SURF algorithm in FPGA: A novel architecture for high demanding industrial applications. In *2012 Design, Automation Test in Europe Conference Exhibition (DATE)*, pages 161–162, March 2012.
- [2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *Computer vision-ECCV 2006*, pages 404–417, 2006.
- [3] J. Blinn. Consider the lowly 2 x 2 matrix. *IEEE Computer Graphics and Applications*, 16(2):82–88, March 1996.
- [4] P. Burt and E. Adelson. The Laplacian Pyramid as a Compact Image Code. *IEEE Transactions on Communications*, 31(4):532–540, April 1983.
- [5] S. Chappell, A. Macarthur, D. Preston, D. Olmstead, B. Flint, and C. Sullivan. Exploiting Real-time FPGA Based Adaptive Systems Technology for Real-time Sensor Fusion in Next Generation Automotive Safety Systems. In *The IEE Seminar on Target Tracking: Algorithms and Applications 2006 (Ref. No. 2006/11359)*, pages 61–68, March 2006.
- [6] W. Chen, S. Ding, Z. Chai, D. He, W. Zhang, G. Zhang, Q. Peng, and W. Luo. FPGA-Based Parallel Implementation of SURF Algorithm. In *2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS)*, pages 308–315, December 2016.
- [7] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and Vision Computing*, 10(3):145–155, April 1992.
- [8] D. Chetverikov, D. Svirko, D. Stepanov, and P. Krsek. The Trimmed Iterative Closest Point algorithm. In *Object recognition supported by user interaction for service robots*, volume 3, pages 545–548 vol.3, 2002.
- [9] L. M. Ledesma-Carrillo, E. Cabal-Yepez, R. d J. Romero-Troncoso, A. Garcia-Perez, R. A. Osornio-Rios, and T. D. Carozzi. Reconfigurable FPGA-Based Unit for Singular Value Decomposition of Large m x n Matrices. In *2011 International Conference on Reconfigurable Computing and FPGAs*, pages 345–350, November 2011.
- [10] Hui Li, B. S. Manjunath, and S. K. Mitra. Multi-sensor image fusion using the wavelet transform. In *Proceedings of 1st International Conference on Image Processing*, volume 1, pages 51–55 vol.1, November 1994.
- [11] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
- [12] Wencheng Wang and Faliang Chang. A Multi-focus Image Fusion Method Based on Laplacian Pyramid. *Journal of Computers*, 6(12), December 2011.
- [13] Y. Wang, K. Cunningham, P. Nagvajara, and J. Johnson. Singular Value Decomposition Hardware for MIMO: State of the Art and Custom Design. In *2010 International Conference on Reconfigurable Computing and FPGAs*, pages 400–405, December 2010.