Least Squares Congealing for Unsupervised Alignment of Images

Mark Cox and Sridha Sridharan
Queensland University of Technology
Brisbane, QLD 4001, Australia
{md.cox,s.sridharan}@qut.edu.au

Simon Lucey and Jeffrey Cohn
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{slucey,jeffcohn}@cs.cmu.edu

Abstract

In this paper, we present an approach we refer to as "least squares congealing" which provides a solution to the problem of aligning an ensemble of images in an unsupervised manner. Our approach circumvents many of the limitations existing in the canonical "congealing" algorithm. Specifically, we present an algorithm that:- (i) is able to simultaneously, rather than sequentially, estimate warp parameter updates, (ii) exhibits fast convergence and (iii) requires no pre-defined step size. We present alignment results which show an improvement in performance for the removal of unwanted spatial variation when compared with the related work of Learned-Miller on two datasets, the MNIST hand written digit database and the MultiPIE face database.

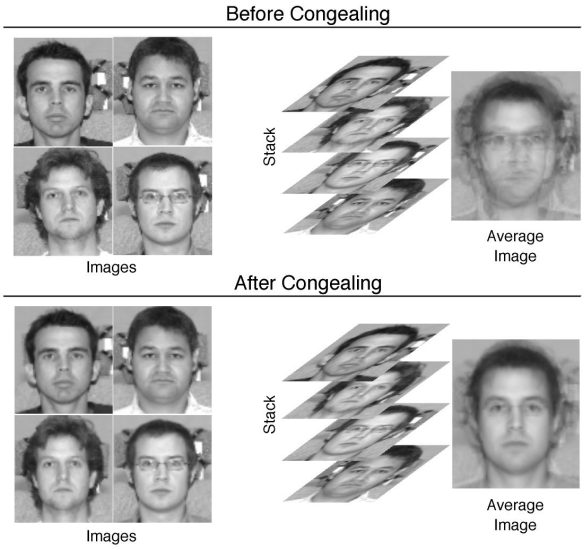


Figure 1. The removal of spatial variations introduced by scale, rotation or translation is an important component of many computer vision systems. Here, four images of faces have their spatial variations removed using the unsupervised process of congealing. The effects of the operation can be seen in the level of detail now present in the average image.

1. Introduction

The task we address in this paper is the automatic alignment of an ensemble of misaligned images in an unsupervised manner. Most recently, this task has been referred to as "congealing" based on the seminal work of Learned-Miller [7] [1]. The only assumption we make in congealing is that the parametric nature of the misalignment is known a priori (e.g. translation, similarity, affine, etc.) and that the images in the ensemble have similar appearance when aligned (e.g., faces, cars, digits, etc.).

It is clear that the capability to congeal an ensemble of misaligned images stemming from the same object class (see Figure 1) has numerous applications in object recognition, detection and tracking. An example of its inclusion in a larger system for the purpose of recognizing objects can be found in [6]. There are a number of issues, however, that are stifling the effective employment of congealing within the wider computer vision and machine learning community.

Specifically, Learned-Miller's approach to congealing employs a sum of entropies cost function to minimize the parametric warp differences between an ensemble of images. This entropy cost function, however, is based on discrete pixel histograms and as a result exhibits poor characteristics when employing common non-linear optimization approaches. To counter this, he had to employ an ad-hoc optimization strat-

[1] Although the original congealing approach could equally be applied to spatial and intensity misalignments, the work conducted in this paper shall concentrate solely on the spatial aspects of this work.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

egy based on a sequential warp parameter update with fixed step sizes. This optimization approach, although receiving excellent performance in [7], is extremely sensitive to the order in which this sequential optimization is conducted and the step size used (see Section 2.2). In its current form the application of the congealing method to different parametric warps and different object classes is non-trivial. A more in depth overview of these issues is given in Section 2.

The work presented in this paper attempts to alleviate many of these problems. We make the following contributions:

• Propose the employment of a sum of squared differences (SSD) cost function for the congealing operation. The employment of this cost function allows for the effective application of a Gauss-Newton gradient descent approach that is: (i) able to simultaneously estimate the warp parameter update, (ii) exhibit fast convergence and (iii) require no pre-defined step size (Section 3).

• Demonstrate that when Gauss-Newton optimization is employed with the SSD cost function our approach is similar to the Lucas & Kanade method to image alignment. However, our proposed approach extends the Lucas & Kanade approach to deal with the alignment of an ensemble of images rather than a single image (Section 3).

• Finally, we demonstrate superior automatic alignment performance, with respect to Learned-Miller's approach, on the MNIST hand written digit database (Section 4.1) and the MultiPIE face database [8] (Section 4.2).

1.1. Related Work

Apart from the work of Learned-Miller, which is of central focus in this paper, there exists a large body of work on the automatic alignment of an image ensemble. Much of this previous work, however, has centered around extending principal component analysis (PCA) to handle the effects of spatial variations. Notably, Frey and Jojic [4] proposed a method for obtaining a set of automatically aligned basis images using the EM algorithm. The approach employed discrete hidden variables to model unwanted spatial variation. A major drawback to this approach, however, was the need to define a discrete set of allowable spatial warps. Additionally, the size of this set directly affected computation time.

Extensions on Frey and Jojic's approach have been proposed by Schweitzer [10] and De la Torre [3]. In these extensions both authors frame the problem of estimating a set of automatically aligned basis images as a bi-linear optimization problem. An advantage of both these approaches is that the spatial warp variation is now modeled continuously rather than discretely. De la Torre improves the situation further by employing additional techniques like genetic algorithms and coarse-to-fine gradient descent techniques to solve the bilinear model. However, the iterative algorithm used to solve these approaches requires estimates of the basis images. In the work of Schweitzer, the estimates are calculated from the initial set of unaligned images. This results in an algorithm which is susceptible to local minima as the level of spatial warp variation governs the quality of the initial estimates of the basis images [10]. For the work of De la Torre, the basis images are initialized from starting frames of video sequences where the motion between frames is assumed to be small. This limits the algorithm to dealing with sequences of video rather than an arbitrary ensemble of images and as a result, does not have to deal with the same amount of appearance variation.

2. Congealing

Congealing can be defined as the minimization of a misalignment function $\mathcal{E}$ which is calculated over a set of N images,

$$\arg\min_{\mathbf{\Phi}} \mathcal{E}(\mathbf{\Phi}) \tag{1}$$

where $\mathbf{\Phi} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{N-1}\}$ is the set of $N-1$ warp parameter vectors corresponding to $N-1$ images in the ensemble and $\mathcal{W}(\mathbf{x};\mathbf{p})$ is the parametric warp function for the pixel coordinates $\mathbf{x}$ in each image. Only $N-1$ image warp parameter vectors need to be found in this optimization as we are trying to automatically align images in the ensemble with each other, so one image needs to remain static. The choice of which image remains static is arbitrary.

Irrespective of the choice $\mathcal{E}()$, minimizing Equation 1 is a highly non-linear and computationally costly operation. Recently, Learned-Miller [7] provided a simplification to this optimization problem. The strategy involved iteratively aligning a single held out image I_i from the rest of the ensemble, or as he referred to it, *stack*. The set of parameters $\mathbf{\Phi}$ were then obtained by sequentially aligning and then updating each image in the stack until $\mathcal{E}()$ converges. An outline of this process can be seen in Figure 2.

The vector function $\mathrm{I}(\mathbf{p}) = [I(\mathcal{W}(\mathbf{x}_1;\mathbf{p})), \ldots, I(\mathcal{W}(\mathbf{x}_M;\mathbf{p}))]^T$ refers to the warped image vector of M intensity pixels. The image $\mathrm{I}(\mathbf{0})$ in Equation 3 refers to the image obtained when applying the identity warp $\mathbf{p} = [0, 0, ..., 0]^T$ (i.e., no warp displacement). For convenience we shall denote $\mathrm{I}(\mathbf{0})$ as I. The image I_i refers to the ith image in the stack.

repeat
  for $i = 1$ to $N - 1$ **do**

$$\mathbf{p}_i \leftarrow \arg\min_{\mathbf{p}} \mathcal{E}_i(\mathbf{p})) \quad (2)$$

$$\mathrm{I}_i(\mathbf{0}) \leftarrow \mathrm{I}_i(\mathbf{p}_i) \quad (3)$$

  end for
until $\mathcal{E}()$ has converged

Figure 2. The iterative congealing algorithm.

Equation 2 employs a new misalignment function which is dependent on only a single warp parameter vector rather than the $N - 1$ warp parameter vectors in Equation 1. As a result, this new misalignment function can be solved in a far more efficient manner depending on the type of measure employed in the misalignment function.

2.1. Learned-Miller Congealing

The measure of misalignment $\mathcal{E}_i()$ used by Learned-Miller in Equation 2 is constructed as a sum-of-entropies function and is shown in Equation 4

$$\mathcal{E}_i(\mathbf{p}) = -\sum_{\mathbf{x}} \sum_{k \in K} p_{\mathbf{x}}(k) \log(p_{\mathbf{x}}(k)) \quad (4)$$

where K is the set of intensity values that a single pixel is allowed to take and $p_{\mathbf{x}}(k)$ is the probability of the pixel intensity at an image coordinate $\mathbf{x}$. This probability is calculated using a histogram of the set,

$$[I_1(\mathbf{x}), \ldots, I_i(\mathcal{W}(\mathbf{x}; \mathbf{p})), \ldots, I_N(\mathbf{x})]^T \quad (5)$$

which consists of the pixel intensities at the same coordinate $\mathbf{x}$ of each image with the ith image warped using the parameter vector $\mathbf{p}$. The rational for using such an approach is that the total entropy for an ensemble of aligned images is less than the total entropy in an ensemble of misaligned images.

Learned-Miller solves for $\mathbf{p}_i$ in Equation 2 by incrementing or decrementing single elements of $\mathbf{p}_i$ by a user defined amount. If the change does not cause an improvement, the value is reset to its initial value and the next element is searched. The amount each element is adjusted by is represented by the step size vector $\mathbf{\Delta} = \{\delta_1, \delta_2, \ldots, \delta_P\}$ and must be defined beforehand.

2.2. Learned-Miller Optimization

The motivation in [7] to use sequential improvement of a single parameter by evaluation instead of a more sophisticated solution, which simultaneously obtains the parameter vector values directly from the measure of misalignment, stems from the problem that the entropy surface is not a smooth function of the affine warp parameters [7].[2] Whilst the sequential algorithm obviously works, the approach incurs non-ideal characteristics. This is best illustrated by the way [7] handles rotation.

Rotation is typically dependent on four of the six parameters in the canonical parametric vector representation of an affine warp. Obviously, this cannot be accounted for when sequentially improving single parameters. Learned-Miller solved this drawback by incorporating an additional redundant parameter which maps to the four fundamental parameters. Whilst this solution is simple for the intuitive affine warp, the number of redundant parameters that may be required for more complicated types of functions like piecewise affine warps could be substantial. In addition, whilst a mathematical definition of the unwanted variation may be specified, common occurrences of variations which are combinations of parameters may be unknown or difficult to determine.

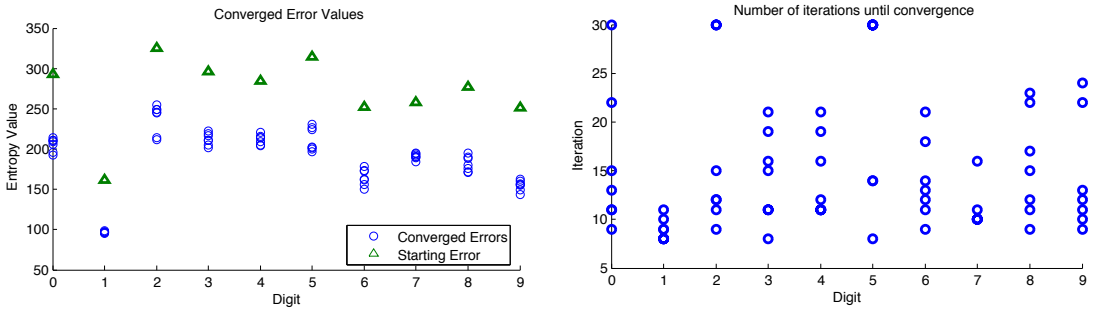
Another aspect that is not ideal in this optimization strategy is the need to specify a step size vector $\mathbf{\Delta}$. Intuitively, if large values are used it will miss subtle variations and the reverse argument applies for small values. This problem could be solved by performing multiple passes of the algorithm in a coarse to fine manner, but for the case where larger numbers of parameters for warp $\mathcal{W}$ are required, iteratively discovering the best values for the step size vector for each pass is suboptimal.

A deeper problem also occurs when attempting to determine the order in which the parameters are to be sequentially improved. To investigate this problem, an experiment was conducted in which the first 300 samples of handwritten digits from LeCun's and Corte's MNIST database [8] was congealed using 7 different parameter orders.[3] For each different parameter order, the resultant entropy at convergence and the number of iterations to attain convergence was collected. This data is shown in Figures 3(a) and 3(b) and shows that the parameter order does effect the performance of Learned-Miller's approach.

Parameter Drift: Empirical observations of the algorithm by Learned-Miller reveal that the algorithm suffers from an average parameter drift during the alignment process. To combat this, he proposed an amendment to the algorithm which removes the average parameter drift by simply removing the average change in

[2] A simultaneous solution was used successfully employed by Learned-Miller to remove variations in brightness in magnetic resonance images of the human brain.

[3] The variations were formed by simply cycle shifting the start ordering of translation: x then y, rotation, scale: x then y and then shear: x then y.



(a) Converged error values

(b) Number of iterations until convergence

Figure 3. The use of a sequential improvement of parameters optimization strategy for minimizing the measure of misalignment introduces some non-ideal characteristics to the congealing operation. One of which is shown here where we show the effect of varying the parameter order in which parameters are improved on (a) the final converged measure of misalignment and (b) the number of iterations to reach the converged value.

the parameters for each image. This approach requires the parameter values and importantly the redundant parameters to be additive, which places additional restrictions on the formulation of the parametric function $\mathcal{W}()$.

3. Least Squares Congealing

This section outlines an alternative method for aligning images which obtains all elements of parameter vector $\mathbf{p}$ simultaneously rather than employing the sequential method presented in Section 2. Our approach is made possible by using an alternate measure of misalignment $\mathcal{E}_i(\mathbf{p})$ which results in an error surface suitable for gradient based optimization techniques. Our approach is motivated by the classic Lucas & Kanade algorithm [9] for iteratively aligning a single image with respect to another using gradient-descent optimization. In our proposed approach we employ a SSD function for misalignment,

$$\mathcal{E}_i(\mathbf{p}) = \sum_{\substack{j=1 \\ j \neq i}}^{N} \left[\mathrm{I}_j - \mathrm{I}_i(\mathbf{p}) \right]^2 \tag{6}$$

As it stands, Equation 6 is still a highly non-linear function and still difficult to minimize. We can linearize this equation by taking the first order Taylor series approximation around $\mathrm{I}_i(\mathbf{p})$ where $\mathbf{p}$ is now our initial guess of the true alignment and $\Delta\mathbf{p}$ is what we are now trying to explicitly estimate,

$$\arg\min_{\Delta\mathbf{p}} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left[\mathrm{I}_j - \mathrm{I}_i(\mathbf{p}) - \frac{\partial \mathrm{I}_i(\mathbf{p})}{\partial \mathbf{p}}^T \Delta\mathbf{p} \right]^2 \tag{7}$$

where $\frac{\partial \mathrm{I}_i(\mathbf{p})}{\partial \mathbf{p}}$ are the steepest descent images formed using $\frac{\partial \mathrm{I}_i(\mathbf{p})}{\partial \mathbf{p}} = \frac{\partial \mathcal{W}}{\partial \mathbf{p}} \nabla \mathrm{I}_i(\mathbf{p})$. The solution to Equation 7 is given by,

$$\Delta\mathbf{p} = \mathbf{H}_a^{-1} \frac{\partial \mathrm{I}_i(\mathbf{p})}{\partial \mathbf{p}} \left[\left(\frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \mathrm{I}_j \right) - \mathrm{I}_i(\mathbf{p}) \right] \tag{8}$$

where,

$$\mathbf{H}_a = \frac{\partial \mathrm{I}_i(\mathbf{p})}{\partial \mathbf{p}} \frac{\partial \mathrm{I}_i(\mathbf{p})}{\partial \mathbf{p}}^T \tag{9}$$

we refer to $\mathbf{H}_a$ as the pseudo-Hessian. An iterative solution to Equation 6 can now be found by iteratively solving for $\Delta\mathbf{p}$ and updating the initial guess $\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p}$ until convergence. This type of optimization is typically referred to as *Gauss-Newton* optimization [2].

What becomes immediately obvious, from inspecting Equation 8, is that the incremental update $\Delta\mathbf{p}$ is not estimated from the stack of images, but is merely estimated from the "average" image of all I_js in the stack. In earlier work, Schweitzer [10] comments on this problem by mentioning that the principal components of unaligned images are dominated by variations in scale, rotation and translation rather than appearance, resulting in blurry basis images. Thus, if the basis images are blurry, then the average image is sure to be blurry. By using the average image to control the direction of $\Delta\mathbf{p}$, with all the fine detail of the object lost, the algorithm is again at the mercy of the initial conditions.

In our approach we propose a novel way to circumvent this limitation where we invert the problem being solved in Equation 7,

$$\arg\min_{\Delta\mathbf{p}} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left[\mathrm{I}_j(\mathbf{p}) + \frac{\partial \mathrm{I}_j(\mathbf{p})}{\partial \mathbf{p}}^T \Delta\mathbf{p} - \mathrm{I}_i \right]^2 \tag{10}$$

so that we can solve for $\Delta\mathbf{p}$ by,

$$\Delta\mathbf{p} = \mathbf{H}_b^{-1}\left[\sum_{\substack{j=1\\j\neq i}}^{N}\frac{\partial \mathrm{I}_j(\mathbf{p})}{\partial\mathbf{p}}\Big(\mathrm{I}_j(\mathbf{p}) - \mathrm{I}_i\Big)\right] \tag{11}$$

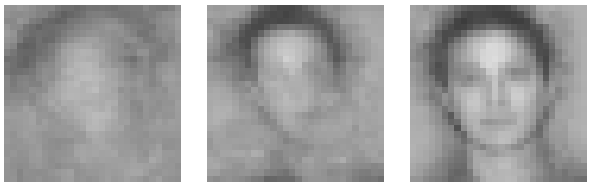
and the new pseudo-Hessian is defined as,

$$\mathbf{H}_b = \sum_{\substack{j=1\\j\neq i}}^{N}\frac{\partial \mathrm{I}_j(\mathbf{p})}{\partial\mathbf{p}}\frac{\partial \mathrm{I}_j(\mathbf{p})}{\partial\mathbf{p}}^T \tag{12}$$

By inverting the problem we are now attempting to estimate the $\Delta\mathbf{p}$ that best aligns the stack to the left out image, rather than the other way around. A major advantage of our proposed solution lies in its ability to use more of the details of each image in the stack for alignment, rather than just relying on the average image of the stack. Inspecting Equation 11 we can see that the update is now being estimated from the steepest descent images stemming from the entire stack rather than a single steepest descent image as in Equation 8. Similarly, the pseudo-Hessian being employed in Equation 12 is based on all the steepest descent images in the stack rather than a single steepest descent image as in Equation 9.

Congealing Using the Average Image: To show the suboptimal effects of using the average image for controlling the congealing process, we setup a simple experiment involving 30 random images of cropped faces from the MultiPIE dataset. We further exaggerated the spatial variations present in the images by randomly perturbing these images in order to remove all detail of the face from the average image (left most image of Figure 4). The randomly perturbed images were then congealed using the method requiring the average image as per Equation 8, and the inverted least squares congealing formulation shown in Equation 11. After each method had performed 10 iterations of congealing, the average image of the final iteration was obtained. The average images for both algorithms can be seen in Figure 4 where the middle and right images correspond to congealing using the average image for alignment and the inverted least squares algorithm respectively. As can be seen, the output of the inverted least squares congealing algorithm is clearly of a face with some detail of the mouth, eye, nose and chin; and as predicted, using the average image for alignment struggles to align the images.

To summarize, the important differentiating features between the least squares congealing algorithm and the Learned-Miller algorithm are: (i) the algorithm uses a different measure of misalignment which has been successfully used for aligning two images and is suitable for gradient descent based techniques [9], [1] (ii) the parameters $\mathbf{p}$ are obtained simultaneously and directly from the measure of misalignment through Gauss-Newton gradient-descent and (iii) no step size set $\mathbf{\Delta}$ is required.



(a) (b) (c)

Figure 4. Using the average image for alignment results in an algorithm which is entirely dependent on the initial conditions of the average image. The proposed algorithm avoids the use of the average image resulting in increased performance for severely misaligned images. This can be seen from the average images of: (a) original data, (b) output of the congealing algorithm where the average image is used for controlling the alignment and (c) output of the inverted least squares congealing algorithm.

3.1. Robust Least Squares Congealing

Robust error functions have been used to great effect in limiting the effects of outliers within gradient descent image alignment [1]. For added robustness we have incorporated this framework into our least-squares congealing framework. A reformulation of Equation 6 can be seen with the robust error function $\varrho()$,

$$\mathcal{E}_i(\mathbf{p}) = \sum_{\substack{j=1\\j\neq i}}^{N}\sum_{\mathbf{x}}\varrho\Big(\left[I_j(\mathcal{W}(\mathbf{x};\mathbf{p})) - I_i(\mathbf{x})\right]^2;\phi\Big) \tag{13}$$

Unfortunately, in the process of introducing the robust error function, an additional parameter ϕ and an additional function $\varrho()$ are required. Determining the function and the parameters are obviously data specific and for this situation where alignment is carried out iteratively on each image, it is likely that the parameters to $\varrho()$ are specific to the held out image and possibly to the current iteration as well, rather than fixed for all iterations. However, for the sake of simplicity, we assume that the parameter ϕ is fixed for all iterations.

An exponential distribution (Equation 14) was chosen for $\varrho()$ as it approximately models the decaying number of observations for increasing error that the distribution of $(I_j(\mathbf{x}) - I_i(\mathbf{x}))^2$ exhibits. By varying the ϕ parameter it is possible to control the emphasis placed on the small errors relative to the larger errors.

$$\varrho(x;\phi) = \phi e^{-\phi x} \tag{14}$$

Using the same procedure as presented earlier, with an additional Taylor expansion of the robust error function $\varrho()$, the incremental update equation is shown in Equation 15.

$$\Delta \mathbf{p} = \mathbf{H}_{\varrho}^{-1} \left[\sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{\partial \mathrm{I}_j(\mathbf{p})}{\partial \mathbf{p}} \mathbf{R}_j \Big(\mathrm{I}_j(\mathbf{p}) - \mathrm{I}_i \Big) \right] \tag{15}$$

$$\mathbf{H}_{\varrho} = \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{\partial \mathrm{I}_j(\mathbf{p})}{\partial \mathbf{p}} \mathbf{R}_j \frac{\partial \mathrm{I}_j(\mathbf{p})}{\partial \mathbf{p}}^T \tag{16}$$

The matrix $\mathbf{R}_j$ refers to a square diagonal matrix of dimension $M \times M$ with each value corresponding to the evaluation of the squared pixel error using the first derivative of $\varrho()$.

Parameter Drift: Like congealing, our method also suffers from an average parameter drift. Rather than requiring the parameters to be additive as done in [7], a different approach specific to images was created.

The method works by initializing three points to the top left, top right and bottom right hand corners of each image and tracking their movement through the iterative congealing process. After a single application of congealing has been applied to all images, the parameters to a single warp is calculated such that when it is applied to the tracked points, it results in the average of the points being their initialized positions.

4. Performance Evaluation

4.1. Congealing Handwritten Digits

In order to compare the two congealing algorithms we setup an experiment in which we attempted to remove the spatial variation present in samples of handwritten digits from the MNIST Dataset [8] using congealing. A total of 50 random samples of each digit were selected from the database of which the average image of those samples can be seen in the first row of Figure 5. The smudged look present in these images indicates that there is considerable spatial variation between samples.

The 50 samples were then congealed using the Learned-Miller[4] and the proposed least squares congealing algorithm. The average image of their output can be seen in the middle and bottom rows of Figure 5 respectively. The sharpness of the average digits of the least squares congealing output shown in Figure 5 compared to the original average digit indicates that the proposed algorithm is an effective method for performing auto-alignment on the handwritten digit samples. The fact that the least squares algorithm does not require step sizes (i.e. $\mathbf{\Delta}$) produces an even cleaner alignment than the Learned-Miller algorithm, as the small spatial variations which cause the blurriness about the edges of the Learned-Miller output have been removed.

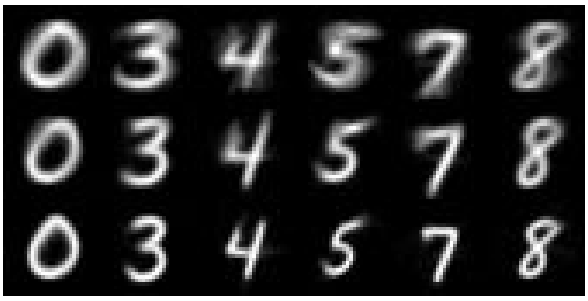


Figure 5. The smudged look of the average image of a set of images is a good indicator of the presence of spatial variations across images. Congealing offers an unsupervised method of removing this unwanted variation. Here, the average image of 50 sample images for a number of handwritten digits is shown: before unsupervised alignment using congealing (top), after congealing using the Learned-Miller algorithm (middle), and after congealing using the proposed algorithm - least squares congealing (bottom).

In order to produce the above results, both methods performed a maximum of 25 iterations of congealing. For the Learned-Miller algorithm, the values which form the step size vector $\mathbf{\Delta}$ were obtained empirically through a number of trials and selecting the configuration which produced the best looking results. The maximum number of incremental update calculations for the calculation of $\mathbf{p}$ in the least squares algorithm was clamped to 25. In the event that $\mathbf{p}$ had not converged by 25 iterations, the image was left unchanged. This condition avoids the case where the image is so severely transformed that the resulting image becomes unidentifiable.

Of interest in the above comparison, is the difference in the number of iterations required for convergence. To determine this, the sum-of-entropies measure of misalignment used in the Learned-Miller method was applied to the output of each iteration of both least squares algorithms. Figure 6 shows the output of this measure after each iteration for the digits three and seven. What is immediately apparent is that the converged point for the least squares congealing algorithm is attained very quickly due to the simultaneous discovery of all parameters.

[4]The Author's own implementation of the Learned-Miller algorithm

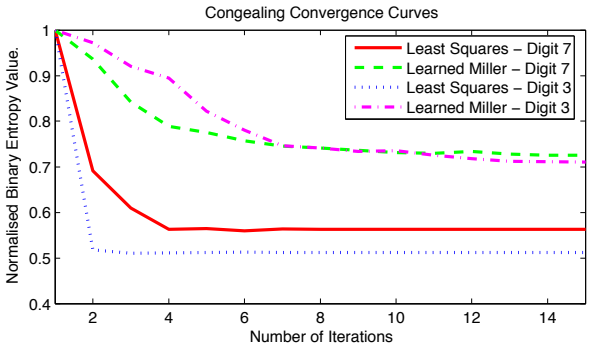


Figure 6. Both the Learned-Miller and least squares congealing algorithms are iterative methods of removing unwanted spatial variations from images. This figure illustrates the number of iterations required for both algorithms to converge to a point where the algorithm can no longer remove any more spatial variation present in 50 images of the handwritten digits '3' and '7'.

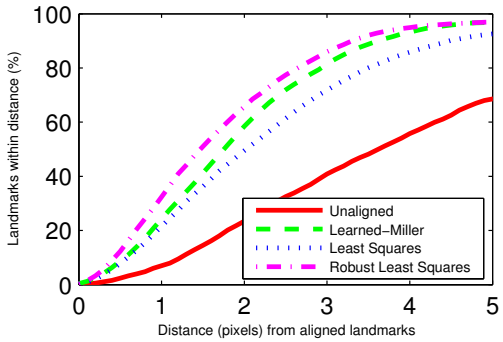


Figure 7. The cumulative distribution of the RMS point error calculated from the difference between a set of 68 annotated landmarks for each face aligned manually and the locations of the same landmarks after congealing.

4.2. Congealing Gray-Scale Cropped Faces

Motivated by the outcome on the MNIST data set, the authors pursued a formal evaluation on a more difficult task: the alignment of faces from the MultiPIE data set [5]. The manually annotated landmarks available with the MultiPIE data set provide a good source of ground truth data that can be used to easily align the images, and they can also be tracked from their initial position throughout the congealing process.

The unaligned reference set of landmarks refers to the original manually annotated landmarks. There are a total of 68 landmarks for each face where each landmark corresponds to an important structural position of the face. The unaligned reference set should ideally be the upper limit of any congealing algorithm. A ground truth aligned set of landmarks was created from the unaligned reference landmarks by aligning all of the landmarks to a single held out image using the similarity transform (scale, translation and rotation).

The position of the landmarks for the output of both congealing algorithms was calculated by applying the resultant affine warp $\mathbf{p}_i$ for each image to the initial set of landmarks for the image. In order to compute a metric of performance for each of the algorithms, the distance between the tracked landmark from congealing and its position in the aligned set was used. Prior to calculating this value, a similarity transform was calculated between the average landmark positions of each algorithm and the aligned set. This transform was then applied to all images in the congealed landmark set to remove any global differences in scale, rotation or translation.

A subset of 30 images were selected where each image was captured under uniform lighting conditions and displaying a neutral facial expression. The images of the face were first cropped to a point where the entire average face was contained in the cropped area with some margin for outliers. Once cropped, the images were interpolated to a size of 28×28 pixels.

The results shown in Figure 7 illustrate the cumulative distribution of the distances between the landmarks in the aligned set and the position of the landmarks after congealing. The result of an ideal algorithm in Figure 7 is to obtain 100% of the landmarks within the smallest distance as possible. The figure shows that an improvement in performance over the Learned-Miller algorithm by using the robust least squares congealing algorithm. Figure 8 shows a portion of the output of the robust least squares congealing operation presented in Figure 7.

Another variable to consider with congealing is what effect does the image size have on performance. In this experiment, the images are rescaled square images of varying side lengths. Results from this experiment indicate similar levels of performance between the Learned-Miller algorithm and the robust least squares algorithm, with each algorithm exhibiting stable behavior around an average RMS point error of 2.25 for images with a side length of 30, 40, 50, 60, and 70 pixels.

5. Discussion and Conclusion

This paper presents an extension to the canonical "congealing" algorithm of Learned-Miller. Our approach, which we refer to as "least squares congealing" has a number of advantages over conventional congealing. Specifically, it: (i) is able to simultaneously, rather than sequentially, estimate warp parameter updates, (ii) exhibits fast convergence and (iii) requires no pre-

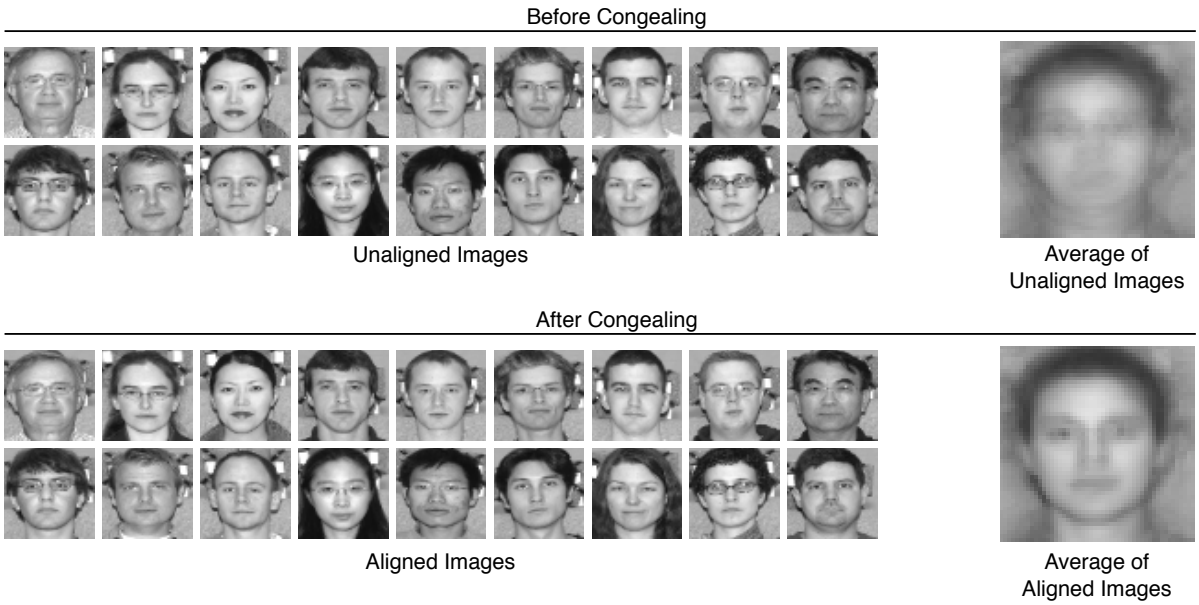


Figure 8. An ensemble of images have their spatial variation removed in an unsupervised manner using the proposed algorithm "least squares congealing". The images used are from the Multiple Pose, Illumination and Expression (MultiPIE) database.

defined step size. A further advantage of our approach is that it now makes the experimentation of congealing on different types of objects and warps trivial as much of the guess work concerning: (i) the order of sequential parameter optimization, (ii) the parametric form of the warp, (iii) the step size to choose; has been eliminated. From an implementation perspective, our proposed algorithm has an additional benefit as it is a straight forward extension of the traditional Lucas & Kanade approach to image alignment.

One drawback to our current approach is that the least squares congealing method is reasonably computationally intensive. Future work shall try and greatly expand our work to develop congealing algorithms that can efficiently align image ensembles containing thousands if not millions of images. Other work shall also attempt to see if congealing can be applied to aligning video sequences of an object in an unsupervised manner and also improve conventional methods for object tracking.

References

[1] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework: Part 1,2 & 3. Technical Report CMU-RI-TR-02-16, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, July 2002.

[2] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, Feb. 2004.

[3] F. De la Torre and M. J. Black. Robust parameterized component analysis: theory and applications to 2D facial appearance models. *Computer Vision and Image Understanding*, 91(1-2):53–71, 2003.

[4] B. Frey and N. Jojic. Transformed component analysis: joint estimation of spatial transformations and image components. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1190–1196 vol.2, 1999.

[5] R. Gross, I. M. S. Baker, and T. Kanade. The CMU Multiple pose, illumination and expression (Multi-PIE) database. Technical Report CMU-RI-TR-07-08, Robotics Institute, Carnegie Mellon University, 2007.

[6] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *Computer Vision, 2007. Eleventh IEEE International Conference on*, 2007.

[7] E. Learned-Miller. Data driven image models through continuous joint alignment. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(2):236–250, 2006.

[8] Y. LeCun and C. Cortes. The mnist database. Online, May 2007. http://yann.lecun.com/exdb/mnist/.

[9] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 1981 DARPA Image Understanding Workshop*, 1981.

[10] H. Schweitzer. Optimal eigenfeature selection by optimal image registration. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 1, 1999.