Intensity-Based Congealing for Unsupervised Joint Image Alignment

Markus Storer, Martin Urschler, Horst Bischof
Institute for Computer Graphics and Vision, Graz University of Technology
{storer, urschler, bischof}@icg.tugraz.at

Abstract

We present an approach for unsupervised alignment of an ensemble of images called congealing. Our algorithm is based on image registration using the mutual information measure as a cost function. The cost function is optimized by a standard gradient descent method in a multiresolution scheme. As opposed to other congealing methods, which use the SSD measure, the mutual information measure is better suited as a similarity measure for registering images since no prior assumptions on the relation of intensities between images are required. We present alignment results on the MNIST handwritten digit database and on facial images obtained from the CVL database.

1. Introduction

Congealing is the alignment of an ensemble of misaligned images. The only assumption in congealing is the type of geometric misalignment, e.g., translation, similarity, affine, and the assumption of a self-similar appearance class, e.g., faces, cars. There are several applications for congealing, e.g., the registration of a stack of images from different modalities in medical imaging [11] or the alignment of a training database for machine learning algorithms [4].

The seminal work of Learned-Miller [8] termed the notion "congealing". They minimize parametric warp differences between a stack of images by applying a sum of entropies cost function. In recent work of Cox et al. [1] some problems of Learned-Miller are alleviated, namely the slow convergence, the need to select a stepsize and sensitivity on the warp parameterization. In their work, they applied a sum of squared differences (SSD) cost function to allow for an effective application of a Gauss-Newton gradient descent approach. They are able to simultaneously estimate warp parameter updates and they do not need a pre-defined step size as opposed to [8]. Their approach is similar to the well known Lucas & Kanade image alignment with the extension to an ensemble of images rather than a single image. Cox et al. further improved their results for a larger amount of images [2]. In their work, they claim that employing an inverse compositional formulation of least-squares congealing is superior to their additive formulation in [1] and thereby show an increase of alignment performance.

There are some other methods based on subspace techniques for automatically aligning an ensemble of images. Frey and Jojic [3] extended the Principal Component Analysis (PCA) to cope with non-aligned images. They obtained a set of aligned basis images by applying the EM algorithm. However, one major drawback was the need to define a discrete set of allowable spatial warps affecting also computation time. De la Torre and Black [7] and Schweitzer [13] proposed extensions on Frey and Jojic's approach. They learn a subspace, which is invariant to affine or higher order geometric transformations. The advantage is that the spatial warp variation is modeled continuously rather than discretely. The major drawback is the need for estimates of the basis images for the iterative algorithms which limits the applicability of their algorithms.

In this paper we concentrate on the state-of-the-art congealing methods of [1, 2]. They make use of the SSD similarity measure. The SSD measure makes the implicit assumption that the images differ only by Gaussian noise after registration. Only in that case, the SSD measure is optimal [17]. For congealing this is never the case because we have a lot of intraclass variation, e.g., in the case of congealing facial images, different subjects, facial hair, gender or race. In other words, the SSD is not an appropriate cost function for generic image registration, hence we apply a more sophisticated cost function based on the mutual information measure [11, 15, 16]. This information-theoretic criterion is very general and powerful, because it does not depend on any assumption on the data (other than stationarity) and does not assume specific relations between intensities in a pair of images.

Based on the basic image registration method using mutual information we build our congealing approach,

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.364

which is explained in detail in Section 2. Section 3 exhibits experiments congealing handwritten digit images and a stack of facial images. Our approach shows very good congealing results and is furthermore easy to implement. Finally, we discuss and conclude our work in Section 4.

2. Congealing

Congealing in our case is defined as an advanced case of image registration. In image registration we have one image called *moving image* $I_M(\mathbf{x})$ which is deformed to fit the other image, the *fixed image* $I_F(\mathbf{x})$. That is, we have to find a transformation $T_\theta(\mathbf{x})$ that aligns $I_M(\mathbf{x})$ with $I_F(\mathbf{x})$, where θ are the transformation parameters. The optimal transformation is found by minimizing a cost function $\mathcal{C}$ with respect to θ

$$\hat{\theta} = \arg\min_\theta \mathcal{C}\left(\theta; I_F; I_M\right). \tag{1}$$

In the introductory section we noted that the SSD is not an appropriate cost function for generic image registration, hence we apply a more sophisticated cost function based on the mutual information measure [11, 15, 16]:

$$\mathcal{C}\left(\theta; I_F; I_M\right) = \\ -\sum_{m \in L_M} \sum_{f \in L_F} p\left(f, m; \theta\right) \log_2 \left(\frac{p\left(f, m; \theta\right)}{p_F\left(f; \theta\right) p_M\left(m; \theta\right)} \right), \tag{2}$$

where p is the discrete joint probability, p_F and p_M are the marginal probabilities, and L_F and L_M are sets of regularly spaced histogram bins containing intensity values of the fixed and moving image respectively. L_F and L_M together span a 2D joint discrete histogram $h\left(f, m; \theta\right)$ where the joint histogram values are estimated using Parzen windows w_F and w_M representing the fixed and moving image:

$$\begin{aligned} h\left(f, m; \theta\right) = & \frac{1}{\sigma_F \sigma_M} \sum_{\mathbf{x}_i \in \Omega_F} w_F \left(\frac{f - I_F\left(\mathbf{x}_i\right)}{\sigma_F} \right) \\ & \cdot w_M \left(\frac{m - I_M\left(T_\theta\left(\mathbf{x}_i\right)\right)}{\sigma_M} \right). \end{aligned} \tag{3}$$

The scaling constants σ_F and σ_M must equal the intensity histogram bin widths defined by L_F and L_M. These follow directly from the grey-value ranges of I_F and I_M and the userspecified number of histogram bins $|L_F|$ and $|L_M|$.

The joint histogram $h\left(f, m; \theta\right)$ is proportional to the discrete joint probability $p\left(f, m; \theta\right)$ given by

$$p\left(f, m; \theta\right) = \frac{1}{|\Omega_F|} h\left(f, m; \theta\right) \tag{4}$$

where $|\Omega_F|$ is the number of pixels in the fixed image domain Ω_F. The marginal discrete probabilities p_F and p_M of the fixed and moving image are obtained by summing p over m and f, respectively

$$\begin{aligned} p\left(f; \theta\right) &= \sum_{m \in L_M} p\left(f, m; \theta\right) \\ p\left(m; \theta\right) &= \sum_{f \in L_F} p\left(f, m; \theta\right). \end{aligned} \tag{5}$$

The mutual information measure is very general; only a relation between the probability distributions of the intensities of the fixed and moving image is assumed. This cost function is minimized iteratively by a standard gradient descent method [6] in a multiresolution scheme. The advantage of a multiresolution scheme is to start the registration process at lower image complexity to reduce the sensitivity to get stuck in local minima of the cost function. Furthermore the overall runtime is decreased.

In (3) we observe a loop over pixel coordinates $\mathbf{x}_i$ over the fixed image domain Ω_F. In general, it is not necessary to take all coordinates into account, but a smaller amount of coordinates may already suffice [6, 15]. This subsampling strategy leads to a lower computational cost, especially for larger images. We use a random selection of a user-specified number of coordinates $\mathbf{x}_i$. Furthermore the sampling is performed by taking samples off the pixel grid to improve the smoothness of the cost function, as suggested by [10, 14].

The image registration functionality is provided by the *elastix* package [5], which also allows to choose some other cost functions, optimization techniques, subsampling strategies and interpolation methods.

For congealing we extend the concept of image registration. Every image of the ensemble of unaligned images is taken once as a moving image. This happens in an outer loop over all images. During one outer loop iteration all other images serve as fixed images. We register the moving image to every fixed image using the entropy based registration described above obtaining the transformation parameters θ. By averaging the transformation parameters obtained from all those registrations we get the final transformation parameters for the moving image. Note that this procedure is significantly simpler and easier to implement than the comparable approaches of [2, 8].

3. Experimental Results

First we evaluate our congealing algorithm on handwritten digits obtained from the MNIST database [9] in Section 3.1. The outcome of the experiments using handwritten digits motivated us to apply congealing to a

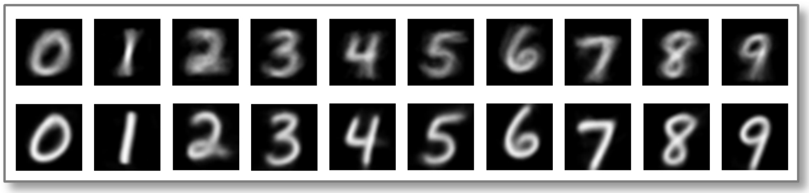

Figure 1. Average images before (first row) and after congealing (second row). The samples were obtained from the MNIST database [9].

more difficult object class. Therefore we show congealing with an ensemble of facial images in Section 3.2.

3.1. Congealing Handwritten Digits

We show the applicability of congealing using samples from the MNIST handwritten digit database [9]. A total of 50 randomly selected images per digit are used for our experiments. We allow an affine transform $T_\theta(\mathbf{x})$ for this image registration task having six parameters to optimize. The results are presented visually in terms of average images. Figure 1 (second row) shows the sharpness of the average images generated from congealing compared to the average images of the unaligned digits in Figure 1 (first row). It can clearly be seen that most of the spatial variation among the digits is removed. The average runtime[1] to congeal one sample of size 28x28 is 78s.

3.2. Congealing Facial Images

Motivated from the results of our congealing experiments with handwritten digits in Section 3.1 we apply our algorithm also to facial images from the CVL database [12], because the images of this database show variation in gender, pose and facial expression and are not aligned. The CVL database consists of facial color images from 114 individuals of a resolution of 640x480 pixels. For our experiments we use the frontal pose image of every individual. We crop the faces to a size of 270x270 pixels and perturb the images randomly by a small amount of translation, scale and rotation to build a strongly unaligned set of facial images. In contrast to our first experiment in Section 3.1 we allow only a similarity transform $T_\theta(\mathbf{x})$ for image registration, because we do not want to shear our facial images. The unaligned facial images and the result of congealing is shown in Figure 4.

To be able to perform also a quantitative evaluation of congealing quality, we annotated all frontal images manually with 19 landmark points at salient facial feature positions, shown in Figure 2. We will provide our annotations for the public research community. We created a set of aligned landmarks used as groundtruth by applying Procrustes Analysis. The Point-to-Point distance of the unaligned landmarks (corresponding to the perturbed images) to the aligned landmarks is illustrated in Figure 3. After congealing we used the obtained landmarks and compared them also to the groundtruth exhibited in Figure 3. It can clearly be seen that the distribution of the Point-to-Point distances is shifted towards smaller displacements emphasizing the applicability of our algorithm. We also want to show the benefits of our mutual information cost function by replacing (2) by a SSD measure and compare the congealing results in Figure 3. The SSD is clearly outperformed by the mutual information measure, especially the SSD exhibits many outliers. These findings substantiate our claims from the introductory section.

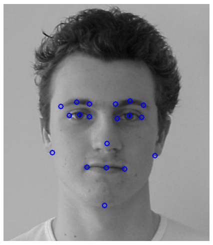

Figure 2. Annotation of a facial image with 19 landmark points at salient facial feature positions.

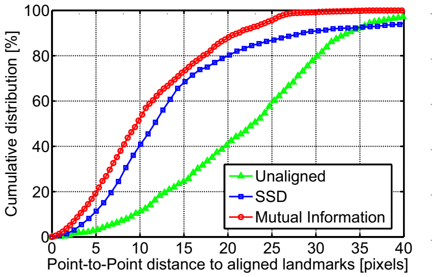

Figure 3. Point-to-Point distance of the unaligned landmarks and the congealed landmarks to the aligned landmarks.

[1] The runtime is measured using an Intel Core 2 Duo processor running at 2.4GHz.

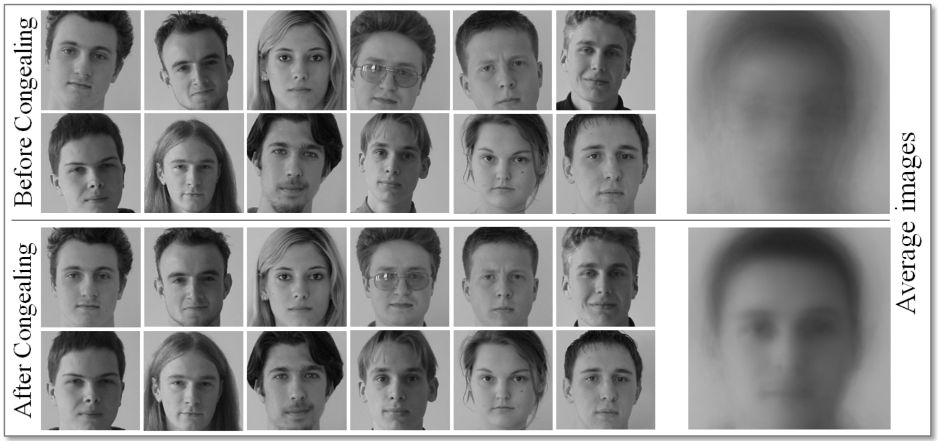

Figure 4. The spatial variation gets removed from the perturbed facial samples by our algorithm. The samples are taken from the CVL database [12].

4. Conclusion

We presented an algorithm for unsupervised alignment of a stack of images. The commonly used SSD measure for congealing is not appropriate for generic image registration. Hence, we used the more sophisticated mutual information similarity measure as a cost function. This cost function is optimized by a standard gradient descent method in a multiresolution scheme. The congealing results on the MNIST handwritten digit database and the results for congealing facial images obtained from the CVL database clearly show the applicability of our algorithm. We also provide our annotations on facial images of the CVL face database for the public research community.

Acknowledgements This work has been funded by the Biometrics Center of Siemens IT Solutions and Services, Siemens Austria.

References

[1] M. Cox, S. Sridharan, S. Lucey, and J. Cohn. Least squares congealing for unsupervised alignment of images. In *Proc. CVPR*, June 2008.

[2] M. Cox, S. Sridharan, S. Lucey, and J. Cohn. Least-squares congealing for large numbers of images. In *Proc. ICCV*, August 2009.

[3] B. J. Frey and N. Jojic. Transformed component analysis: Joint estimation of spatial transformations and image components. In *Proc. ICCV*, volume 2, pages 1190–1196, 1999.

[4] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *Proc. ICCV*, 2007.

[5] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim. elastix: A toolbox for intensity-based medical image registration. *IEEE Trans. on Medical Imaging*, 29(1):196–205, January 2010.

[6] S. Klein, M. Staring, and J. P. Pluim. Evaluation of optimization methods for nonrigid medical image registration using mutual information and b-splines. *IEEE Trans. on Image Processing*, 16(12):2879–2890, 2007.

[7] F. D. 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.

[8] E. G. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE Trans. on PAMI*, 28(2):236–250, 2006.

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

[10] B. Likar and F. Pernus. A hierarchical approach to elastic registration based on mutual information. *Image and Vision Computing*, 19(1-2):33–44, January 2001.

[11] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Trans. on Medical Imaging*, 16(2):187–198, 1997.

[12] P. Peer. CVL face database, Computer Vision Laboratory, University of Ljubljana. http://www.lrv.fri.uni-lj.si/facedb.html.

[13] H. Schweitzer. Optimal eigenfeature selection by optimal image registration. In *Proc. CVPR*, volume 1, 1999.

[14] P. Thevenaz, M. Bierlaire, and M. Unser. Halton sampling for image registration based on mutual information. *Sampling Theory in Signal and Image Processing*, 7(2):141–171, March 2008.

[15] P. Thevenaz and M. Unser. Optimization of mutual information for multiresolution image registration. *IEEE Trans. on Image Processing*, 9(12):2083–2099, 2000.

[16] P. Viola and W. M. Wells III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137154, 1997.

[17] P. A. Viola. *Alignment by Maximization of Mutual Information*. PhD thesis, MIT, 1995.