Image Congealing via Efficient Feature Selection *

Ya Xue
Machine Learning Lab
GE Global Research
xueya@ge.com

Xiaoming Liu
Computer Vision Lab
GE Global Research
liux@resarch.ge.com

Abstract

Congealing for an image ensemble is a joint alignment process to rectify images in the spatial domain such that the aligned images are as similar to each other as possible. Fruitful congealing algorithms were applied to various object classes and medical applications. However, relatively little effort has been taken in the direction of compact and effective feature representations for each image. To remedy this problem, the least-square-based congealing framework is extended by incorporating an unsupervised feature selection algorithm, which substantially removes feature redundancy and leads to a more efficient congealing with even higher accuracy. Furthermore, our novel feature selection algorithm itself is an independent contribution. It is not explicitly linked to the congealing algorithm and can be directly applied to other learning tasks. Extensive experiments are conducted for both the feature selection and congealing algorithms.

1. Introduction

Group-wise image alignment, often coined as *congealing* [28, 23], is defined as a process of jointly estimating warping parameters for all images in an ensemble. There are many applications of image congealing. In the learning of an object detector [40, 12], the position of the object (face/pedestrian/car) for all training images can be automatically provided by congealing, rather than being labeled manually. Congealing is also able to improve appearance-based face recognition performance [19]. Yan *et al.* show that automatic labeling of facial landmarks can be enabled by semi-supervised congealing [37, 26], which can also potentially be used to discover the non-rigid shape deformation of a real-world object.

Congealing aims to estimate the warping parameters by iteratively minimizing a distance metric computed using the feature presentation of each image. Hence, there are three key elements to image congealing: *cost function*, *optimization method*, and *feature representation*. Miller *et al.* [28] utilizes the mutual information as the cost function for optimization, while Cox *et al.* [9] employ a least-squared distance between image pairs in the ensemble. Regarding the optimization method, gradient descent [28] and the inverse compositional approach [2, 9, 37] are all valid choices.

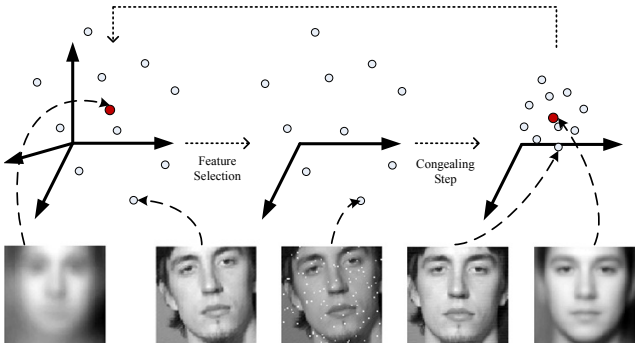


Figure 1. Given an unaligned image ensemble (blurred average image), a novel unsupervised feature selection and congealing are iteratively applied and result in the improved alignment for all images, as indicated by the sharpness in the average image. The white dots on the middle image indicate the selected features.

In contrast, almost all prior work use the original image intensities as the feature representation, which has a number of drawbacks. Since such representation usually resides in a high-dimensional space, it imposes substantial computational burden for optimization, especially with a large image ensemble. Also, because many pixel intensities are redundant due to local proximity to their neighboring pixels, they may hinder the optimization process. To remedy this problem, as shown in Fig. 1, this paper proposes an unsupervised feature selection approach to automatically choose a subset of feature representation and use that for image congealing. We experimentally show that by only using less than 3% of the original feature representation, both the accuracy and efficiency of the congealing can be substantially improved

*Both authors contributed equally to this work. This work was supported by the National Institute of Justice, US Department of Justice, under the award #2009-SQ-B9-K013. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

in comparing to the one without feature selection.

Not only do we marry feature selection to congealing, but we also propose a novel unsupervised feature selection approach. We first construct a graph with features as the vertices and the connectivity between vertices is determined by the *maximum information compression index* [29]. A simple and fast graph clustering method called *power iteration clustering* (PIC) [24] is employed to partition the graph into subsets and select a representative feature from each subset. The proposed method is an extension of the well-known feature clustering algorithm in [29]; nevertheless, our method has significant advantages in efficiency, especially when the feature dimension is high, while achieving comparable effectiveness in terms of removing feature redundancy. Moreover, the proposed method can be easily applied to other learning tasks beyond congealing, due to its independence from the objective function and optimization algorithm for the target concept.

In summary, this paper has two main contributions supported by extensive experiments and comparison with the baseline approaches:

◇ A novel extension to least-square-based congealing based on the finding that incorporating unsupervised feature selection can improve both the accuracy and efficiency of congealing.

◇ A novel unsupervised feature selection algorithm that is simple, fast and generally applicable.

2. Prior Work

There is a long history of group-wise image alignment in computer vision [39, 3, 21, 11, 38, 35, 25, 31], particularly in the area of medical image analysis [8, 4]. Learned-Miller [28, 23] names this process "congealing", where the basic idea is to minimize a cost function by estimating the warping parameters of an ensemble. Over the years, there have been various directions that have been explored to improve the accuracy and efficiency of congealing. In terms of a cost function, the recent work of Storer *et al.* utilizes a mutual information measurement as an objective function [36]. Cox *et al.* [9, 10] and Yan *et al.* [37] develop a series of least-squares-based congealing algorithms. In terms of a learning paradigm, there are unsupervised congealing [3, 9, 23], as well as semi-supervised congealing [37, 26]. The warping function used to compute pair-wise image distances can be defined as a global affine warp [9], or sophistical non-rigid warp [35, 7]. However, one area that has received relatively little attention concerns what is an effective feature representation in the context of congealing. With only a few exceptions, such as the HOG feature in [26], most prior work compute the cost function by directly utilizing the original pixel intensities of the image. Our proposed congealing algorithm makes a sharp contrast in that we develop a novel feature selection mechanism to effectively choose a subset of the feature representation, which is shown to improve both the accuracy and efficiency of least-squares-based congealing.

The task of feature selection is to remove irrelevant and/or redundant features. Irrelevant features refer to the features that are not informative with respect to the target concept (*e.g.*, class in supervised learning); redundant features refer to those that are highly correlated to some other features [41]. By removing the irrelevant and redundant features, feature selection helps reduce over fitting and improve efficiency of model learning. It also helps better understand the underlying data-generating mechanism and related physical process patterns.

Feature selection has been well studied in supervised learning [17]. Nevertheless, far less attention has been paid to feature selection in unsupervised learning, mainly because the definition of relevance becomes unclear without guidance of class labels. A few approaches have been presented in the literature. Following [20], we categorize them into two groups, *wrapper* and *filter*. A wrapper method ties feature selection with the main learning task (*e.g.*, classification) and evaluates features by how well they fit the ultimate learning goal. In contrast, a filter method does not rely on the learning algorithm but exploits intrinsic properties of the data structure.

In the first category, most unsupervised *wrapper* techniques use clustering quality or related constructs as feature selection guidance and are customized to a particular clustering algorithm. Dy and Brodley [14], for example, wrap feature selection around an EM clustering algorithm and measure both the scatter separability and the maximum likelihood. Other examples can be found in [13, 30, 22, 32]. Less techniques have been found in the second category - the *filter* type of unsupervised feature selection techniques. The Laplacian score is proposed in [18] to measure features by their power of locality preserving; Zhao and Liu [42] present a general feature selection framework evolved from the spectral graph theory and shows the Laplacian score algorithm is a special case of the proposed framework. Another work that has received much attention is the feature clustering method presented in [29], which partitions the features into a number of homogenous subsets, according to an information-theory-based similarity measure, and then selects the representative feature for each subset.

For our learning task (congealing), the *filter* techniques are more preferable because clustering is not our ultimate learning objective. Existing filter methods have difficulties with high-dimensional, big datasets, which are quite common in real-world applications. Therefore, we propose a new filter method that is a natural extension of [29] but powered by a fast graph clustering approach. Our method provides a comparable or even better performance of feature selection when independently evaluated on benchmark

datasets. When embedded in the congealing algorithm, its advantage becomes clearer: the optimization search space is shrunk by removing redundant features and therefore the computation cost is reduced by a significant margin.

3. The Congealing Algorithm

First we describe the basic concept and objective function of the conventional unsupervised least-squares-based congealing [9, 37].

Unsupervised congealing approaches operate on an ensemble of K unaligned images $\mathbf{I} = \{\mathbf{I}_i\}_{i=1}^{K}$, each with an unknown warping parameter $\mathbf{p}_i$ that is to be estimated. The $\mathbf{p}_i$ can be a simple 6-dimensional affine warping parameter, or the coefficient parameter of a shape subspace. The goal of congealing is to estimate the collection of all unknown parameters, $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_K]$, by minimizing a cost function defined on the entire ensemble:

$$\varepsilon(\mathbf{P}) = \sum_{i=1}^{K} \varepsilon_i(\mathbf{p}_i). \tag{1}$$

The total cost is the summation of the cost of each image $\varepsilon_i(\mathbf{p}_i)$:

$$\varepsilon_i(\mathbf{p}_i) = \sum_{j=1, j\neq i}^{K} \|f(\mathbf{I}_j, \mathbf{p}_j) - f(\mathbf{I}_i, \mathbf{p}_i)\|^2, \tag{2}$$

where $f(\mathbf{I}, \mathbf{p})$ is a d-dimensional feature representation of image $\mathbf{I}$ evaluated at $\mathbf{p}$. Hence, $\varepsilon_i(\mathbf{p}_i)$ equals the summation of the pairwise feature difference between $\mathbf{I}_i$ and all the other images in the ensemble.

In [37], the feature representation is defined as,

$$f(\mathbf{I}, \mathbf{p}) \doteq \mathbf{I}(\mathbf{W}(\mathbf{x}; \mathbf{p})), \tag{3}$$

where $\mathbf{W}(\mathbf{x}; \mathbf{p})$ is a warping function that takes as input $\mathbf{x}$, which is a collection of all d pixel coordinates within the common rectangle region, and outputs the corresponding pixel coordinates in the coordinate space of image $\mathbf{I}$. Given this warping function, $\mathbf{I}(\mathbf{W}(\mathbf{x}; \mathbf{p}))$ denotes the corresponding warped image feature obtained by bilinear interpolation of the image $\mathbf{I}$ using the warped coordinates $\mathbf{W}(\mathbf{x}; \mathbf{p})$.

Since the total cost $\varepsilon(\mathbf{P})$ is difficult to optimize directly, [37] chooses to iteratively minimize the individual cost $\varepsilon_i(\mathbf{p}_i)$ for each $\mathbf{I}_i$, given an initial estimation of the warping parameter $\mathbf{P}_i^{(0)}$. The well-known inverse warping technique [2] is utilized and after taking the first order Taylor expansion, Eqn. (2) can be simplified to:

$$\sum_{j=1, j\neq i}^{K} \|\mathbf{b}_j + \mathbf{C}_j \Delta\mathbf{p}_i\|^2, \tag{4}$$

where

$$\mathbf{b}_j = f(\mathbf{I}_j, \mathbf{p}_j) - f(\mathbf{I}_i, \mathbf{p}_i), \ \mathbf{C}_j = \frac{\partial f(\mathbf{I}_j, \mathbf{p}_j)}{\partial \mathbf{p}_j}. \tag{5}$$

The least-square solution of Eqn. (4) can be obtained by setting the partial derivative of Eqn. (4) with respect to $\Delta\mathbf{p}_i$ to be equal to zero. We have:

$$\Delta\mathbf{p}_i = -\left[\sum_{j=1, j\neq i}^{K} \mathbf{C}_j^T \mathbf{C}_j\right]^{-1} \left[\sum_{j=1, j\neq i}^{K} \mathbf{C}_j^T \mathbf{b}_j\right]. \tag{6}$$

The calculated $\Delta\mathbf{p}_i$ is used to update the current warping parameter, $\mathbf{p}_i^{(t)}$:

$$\mathbf{p}_i^{(t+1)} \leftarrow \mathbf{p}_i^{(t)} + \Delta\mathbf{p}_i. \tag{7}$$

Similar updating is conducted for the warping parameters of other images in the ensemble, and then the algorithm proceeds to the next iteration. This process terminates when the difference of $\varepsilon(\mathbf{P})$ (computed via Eqn. (1)) between consecutive iterations is less than a pre-defined threshold.

4. Unsupervised Feature Selection

Our feature selection approach is designed to remove feature redundancy. Investigation of feature relevance is beyond the scope of this paper. We aim to develop an unsupervised feature selection algorithm that is suitable for various learning tasks with different target concepts; hence, there doesn't exist a unified definition of feature relevance.

Let $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_d]$ denote a n-by-d data matrix, where rows are instances and columns are features. The vector $\boldsymbol{y}_j$ includes the jth feature for all the instances. Mitra *et al.* [29] propose a feature similarity measure based on information theory termed the *maximum information compression index*, which possesses several desirable properties for feature redundancy reduction, such as sensitivity to scaling and invariance to rotation. It is defined as follows

$$\lambda(\boldsymbol{y}_j, \boldsymbol{y}_l) = \frac{1}{2}\big[V(\boldsymbol{y}_j) + V(\boldsymbol{y}_l) - \sqrt{(V(\boldsymbol{y}_j) + V(\boldsymbol{y}_l))^2 - 4V(\boldsymbol{y}_j)V(\boldsymbol{y}_l)(1 - \tau(\boldsymbol{y}_j, \boldsymbol{y}_l)^2)}\ \big]. \tag{8}$$

where τ is the correlation coefficient, $\tau(\boldsymbol{y}_j, \boldsymbol{y}_l) = \frac{C(\boldsymbol{y}_j, \boldsymbol{y}_l)}{\sqrt{V(\boldsymbol{y}_j)V(\boldsymbol{y}_l)}}$, $V(\cdot)$ the variance of a random variable and $C(\cdot, \cdot)$ the covariance between two variables. The value of λ ranges between 0 and $0.5(V(\boldsymbol{y}_j) + V(\boldsymbol{y}_l))$. It is minimized when two features $\boldsymbol{y}_j$ and $\boldsymbol{y}_l$ are linearly dependent and increases as the dependency diminishes. Based on the measure λ for each pair of features, a heuristic algorithm is employed in [29] to search the feature space: it finds the k nearest neighbors of each feature; the feature with the most compact neighborhood is selected and its neighbors are discarded; the process is repeated until all features are either selected or discarded. The heuristic search algorithm has computational complexity similar to that of a kNN algorithm, which could be quite slow when the feature dimension is high.

Recently, spectral graph theory has become an active research area in machine learning. Spectral clustering algorithms (*e.g.* [34]) could be used to take place of the heuristic search algorithm. If we build a graph $\mathbf{A}$ with features as vertices, the connectivity between vertices can be defined as a function of the measure λ in Eqn. (8)

$$a_{jl} = \exp(-\lambda(\boldsymbol{y}_j, \boldsymbol{y}_l)^2/(2\sigma^2)), \quad j, l = 1, \ldots, d, \quad (9)$$

where σ is a scaling parameter that controls the kernel width [1]. The degree matrix associated with $\mathbf{A}$, denoted by $\mathbf{D}$, is a diagonal matrix with the diagonal entries equal to the row sums of $\mathbf{A}$. A normalized random-walk Laplacian matrix $\mathbf{L}$ is defined as $\mathbf{L} = \mathbf{\Delta} - \mathbf{D}^{-1}\mathbf{A}$ [27], where $\mathbf{\Delta}$ is the identity matrix. The intrinsic clustering structure is often revealed by representing the data in the basis composed of the smallest eigenvectors of $\mathbf{L}$ (but not the very smallest one). The very smallest eigenvector is a constant vector that doesn't have discriminative power.

If we define another matrix $\mathbf{W} = \mathbf{D}^{-1}\mathbf{A}$, its largest eigenvector is the smallest eigenvector of $\mathbf{L}$. A well-known method for computing the largest eigenvector of a matrix is *power iteration* (PI), which randomly initializes a d-dimensional vector $\boldsymbol{v}^{(0)}$ and iteratively updates the vector by multiplying it with $\mathbf{W}$

$$\boldsymbol{v}^{(t)} = \gamma \mathbf{W} \boldsymbol{v}^{(t-1)}, \quad t = 1, 2, \ldots, \quad (10)$$

where γ is a normalizing constant to keep $\boldsymbol{v}^{(t)}$ numerically stable.

Lin and Cohen [24] discover an interesting property of the largest eigenvector of $\mathbf{W}$: before the elements of $\boldsymbol{v}^{(t)}$ converge to the constant value, they first converge to local centers that correspond to the clusters in the data. Therefore, the largest eigenvector $\boldsymbol{v}^{(t)}$, which is discarded in spectral clustering algorithms, becomes a useful tool for clustering. The algorithm, *power iteration clustering* (PIC), is very efficient since it only involves iterative matrix-vector multiplications and clustering the one-dimensional embedding of the original data is a relatively easy task.

In [24] PIC is used to partition the graph with data instances as vertices. Instead, we are interested in feature clustering and PIC is applied to the graph built on features. Once we have the embedding vector $\boldsymbol{v}^{(t)}$, various clustering algorithms can be applied to group the features. To reduce computational cost introduced by the clustering algorithm, we use the fast k-means algorithm presented in [15]. Dirichlet process mixture models [1, 5] could be a solution if the number of clusters, *i.e.* the number of selected features, remains unknown and is considered a model parameter to be estimated as well.

[1] To avoid the issue of parameter selection, the value of σ is automatically set as $\sigma = \text{median}(\{a_{jl}\}_{j,l=1}^{d})$.

Input: data matrix $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_d]$ and number of features to be selected, k (optional)
1. Calculate similarity between every pair of features using Eqn. (8) and Eqn. (9) and build the graph $\mathbf{A}$.
2. Obtain $\mathbf{W}$ by row normalizing $\mathbf{A}$.
3. Initialize $\boldsymbol{v}^{(0)}$ with Eqn. (12).
4. Find the embedding vector with iterative matrix-vector multiplications as in Eqn. (10).
5. Group the elements of the vector, each corresponding to one feature, with an efficient clustering algorithm, *e.g.* the fast k-means [15].
6. Let $\Omega = \emptyset$; in each cluster, find the feature that is closest to the cluster center and assuming its index is j, let $\Omega = \Omega \cup \{j\}$.
Output: indices of selected features, Ω.

Table 1. The proposed unsupervised feature selection algorithm.

There are additional enhancements we have made to the PIC algorithm in order to increase algorithm stability. PIC doesn't work for the following matrix, for example,

$$A = \begin{bmatrix} 0 & 1 & 0.1 & 0 \\ 1 & 0 & 0 & 0.1 \\ 0.1 & 0 & 0 & 1 \\ 0 & 0.1 & 1 & 0 \end{bmatrix}. \quad (11)$$

It is suggested in [24] that initializing $\boldsymbol{v}^{(0)}$ with the degree vector $\boldsymbol{u} = [u_1, u_2, \cdots, u_d]^T$ can accelerate local convergence, where $u_j = \frac{\sum_l a_{jl}}{\sum_{j,l} a_{jl}}$. However, for a matrix like Eqn. (11), the degree vector is a constant vector and will remain constant during the matrix-vector multiplication process. To address this issue and assure fast convergence, we add a small perturbation to the initial vector, *i.e.*

$$v_j^{(0)} = u_j + \varepsilon_j, \quad j = 1, \ldots, d, \quad (12)$$

where ε_j is a small random number, *e.g.* uniformly distributed in the interval $(0, 1e^{-2}/d)$. Then we normalize $\boldsymbol{v}^{(0)}$ to sum one. In addition, we find that setting the diagonal elements of $\mathbf{A}$ to be one (instead of zeros as suggested in [24]) leads to better numerical stability.

Table 1 summarizes the overall procedure of the proposed unsupervised feature selection algorithm.

5. Congealing with Feature Selection

Having introduced the unsupervised feature selection method, we now present how to incorporate it into the unsupervised congealing framework discussed in Section 3.

Given the initial warping parameter $\mathbf{P}^{(0)}$, the basic unsupervised least-square-based congealing algorithm proceeds with the following iterative steps: 1) computing the warping

parameter update $\Delta\mathbf{p}_i$ for each image, and 2) updating the current warping parameter for each image. Our proposed algorithm follows these same steps, except that our feature representation is only a subset of the original presentation, and is defined as,

$$f(\mathbf{I}, \mathbf{p}) \doteq \mathbf{I}(\mathbf{W}(\mathbf{x}(\Omega); \mathbf{p})), \tag{13}$$

where Ω is the output of the unsupervised feature selection method described in Table 1, and is a k-dimensional vector containing the indices of selected features.

There are several aspects regarding this enhanced congealing algorithm. First, although being similar to the case of original representations, the calculation of $\mathbf{b}_j$ and $\mathbf{C}_j$ is more efficient because only the feature elements with indices included in Ω need to be computed. Second, we choose to conduct the unsupervised feature selection at every iteration. The motivation is that as the alignment for all images changes at each iteration, the corresponding visual features also change, which implies that a distinctive subset of features might be useful at different stages of the entire iterations. Third, we utilize the same iteration termination condition as the basic congealing algorithm, where the image difference (see Eqn. (1)) is evaluated using the original feature representation. This is an intuitive choice since different feature selections are conducted at consecutive iterations. Finally, our proposed congealing algorithm is not limited to the feature representation in Eqn. (13), which is an algorithmic choice given the original intensity feature in Eqn. (3). Our feature selection method is applicable to other feature types such as regional histograms.

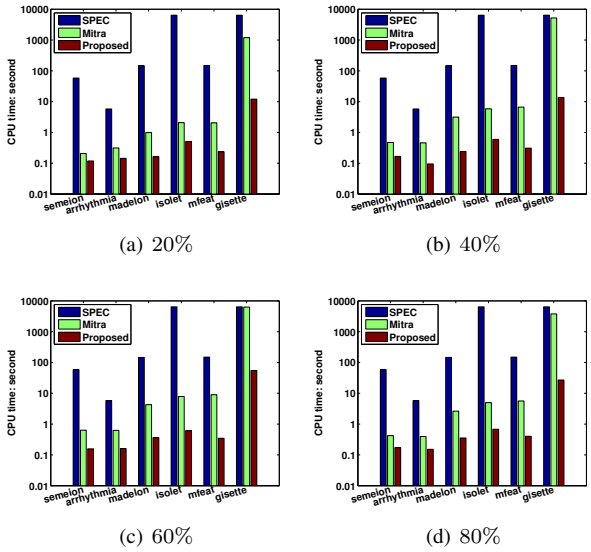
As indicated by [37], the unsupervised least-square-based congealing has computational complexity $\mathcal{O}(mK^2d)$, where m is the dimension of the warping parameter and d is the dimension of the feature representation. Given that the efficiency of congealing depends linearly on the feature dimension, our proposed algorithm has a great potential to improve efficiency by working on a much lower feature dimension k, where $k \ll d$. This is demonstrated by our experiments in Section 6.

6. Experiments

We first compare our proposed feature selection algorithm with state-of-the-art methods. Then we evaluate the unsupervised congealing algorithm with the feature selection. All algorithms are run single threaded on a conventional workstation.

6.1. Evaluation of feature selection performance

We first conduct an empirical study of the proposed feature selection algorithm on several UCI machine learning benchmark datasets, as summarized in Table 2. Following [14, 18, 30, 42], we take a supervised approach to evaluate the quality of selected feature subsets. The ground truth of class labels are inaccessible during the feature selection process and only used to evaluate classification accuracy. The classifier we use is a simple but efficient linear classifier [16], which doesn't have the parameter-tuning issue and has been used for results evaluation in the NIPS 2003 (supervised) feature selection challenge [2].



(a) 20% (b) 40%

(c) 60% (d) 80%

Figure 2. Comparison of CPU time for selecting features with the three unsupervised feature selection methods.

dataset	#(instances)	#(features)	#(classes)
semeion	1593	256	10
arrhythmia	452	279	16
madelon	4400	500	2
isolet	7797	617	26
mfeat	2000	649	10
gisette	13500	5000	2

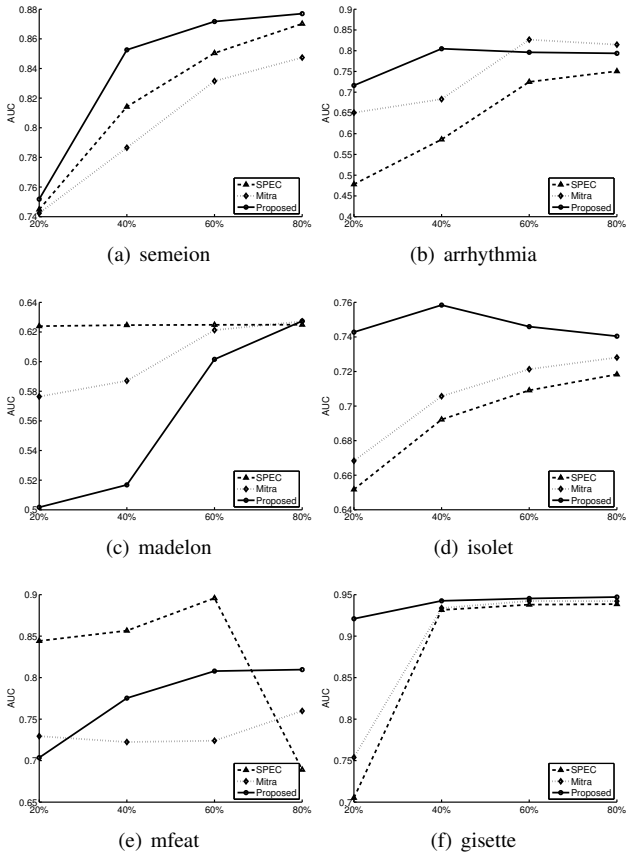
Table 2. UCI datasets used in our experiments, ordered by feature dimension, from low to high.

The performance is evaluated at a different number of selected features, 20%, 40%, 60% and 80% of the original feature dimension. The dataset, with only the selected features, is randomly split into halves, one for training and the other for testing. Classification accuracy is measured by Area Under Curve (AUC), averaged over 100 random splits. If the data includes $M > 2$ classes, the multi-class classification problem is converted into M one-against-all binary classification problems and their average AUC is reported.

We compare three unsupervised feature selection algorithms of the *filter* type, feature clustering in [29] (denoted by "Mitra"), SPEC in [42] and our proposed algorithm. All three algorithms are implemented in non-optimized Matlab™ code [3]. The experiments are run with the default

[2] Information can be found at http://www.nipsfsc.ecs.soton.ac.uk/.

[3] The Mitra algorithm is available at http://www.facweb.iitkgp.ernet.in/∼pabitra/paper.html and the SPEC algorithm at http://featureselection.asu.edu/documentation/spectrum.htm.



(a) semeion (b) arrhythmia

(c) madelon (d) isolet

(e) mfeat (f) gisette

Figure 3. Comparison of feature quality for the three unsupervised feature selection methods, measured by average AUC over 100 random splits.

parameter settings in the original code. To make a fair comparison in efficiency, we use the same code to compute the measure λ in Eqn. (8) for both the Mitra and our algorithm.

The experiments results are reported Figs. 2 and 3, in terms of CPU time and AUC respectively. Our proposed method shows superior efficiency in the comparison of CPU time for feature selection (note that the Y axis in Fig. 2 is in the log scale). It runs less than 1 minute even for a high-dimensional dataset like *gisette* (5,000 features). Taking 20% as example, the CPU times averaged over 6 datasets are 2192, 200, 2 seconds for SPEC, Mitra, and our algorithm respectively.

Classification accuracy for the feature subset selected by our algorithm is comparable to, if not better than, that for the other two algorithms, as shown in Fig. 3. Table 3 shows relative AUC increase averaged over 6 UCI datasets, comparing the proposed algorithm with the Mitra algorithm. The two algorithms only differ in the feature clustering part. Clearly with PIC we improve not only efficiency but also feature selection quality.

It is worth noting that *madelon* is a special dataset in

$k/\#(\text{features})\times 100\%$	20%	40%	60%	80%
$\frac{\text{AUC}_\text{P}-\text{AUC}_\text{M}}{\text{AUC}_\text{M}} \times 100\%$	4.69%	4.99%	2.22%	1.62%

Table 3. Classification improvement over the Mitra algorithm (AUC_M) given by the proposed algorithm (AUC_P), averaged over 6 UCI datasets.

that among its 500 feature, only 20 are real features and all the rest are distracter features having no predicative power. Since it is unknown to us the indices of the real features, we suspect that the SPEC algorithm has the real features ranked among the top 20% and therefore its AUC keeps almost no change as more features are added in. The other two algorithms aim to remove feature *redundancy* and it is likely that they are not able to capture those *relevant* features when the feature grouping is coarse.

6.2. Evaluation of the congealing algorithm

Having demonstrated the effectiveness of our proposed feature selection algorithm, we now focus on its contribution to image congealing.

We collect 300 images from the Notre Dame (ND1) database [6]. We manually label 33 landmarks ($\hat{\mathbf{u}}$) for each image to establish a ground truth and to enable a quantitative evaluation for the congealing performance. During initialization, we add a uniformly distributed random noise $\eta \in [-\eta_{max}, \eta_{max}]$ to the ground-truth value $\hat{\mathbf{u}}_{i,j}$:

$$\mathbf{u}_{i,j} = \hat{\mathbf{u}}_{i,j} + \frac{\eta\rho_i}{\bar{\rho}}, \tag{14}$$

where ρ_i is the eye-to-eye pixel distance of $\mathbf{I}_i$, and $\bar{\rho}$ is the average of ρ_i for all images ($\bar{\rho} \approx 130\ pixels$ in our experiments). By doing so, we may synthesize different levels of deviation in the initialization, which is also relative to the face size. The correspondence between the perturbed landmarks and the average landmarks in the common mean shape are used to generate the initial estimation of warping parameters $\mathbf{P}^{(0)}$ for all images. In practical applications, the initial landmark positions can be obtained from a face detector. A 6-parameter affine warp is employed as $\mathbf{W}(\mathbf{x};\mathbf{p})$. A 72×72 square region is used as the common mean shape in the experiments, which results in a 5184-dimensional representation for the original feature $\mathbf{I}(\mathbf{W}(\mathbf{x};\mathbf{p}))$. Our algorithm is implemented in Matlab™.

The accuracy of the algorithms is evaluated by two criteria: (1) Normalized Root Mean Squared Error (NRMSE) of landmarks defined as the RMSE w.r.t. the ground truth landmarks divided by the eye-to-eye distance ρ_i, and expressed as a percentage; (2) Sample "Outliers" Fraction (SOF) defined as the number of images, of which the NRMSE exceeds a threshold (8%), versus the total number of images. A smaller NRMSE indicates a higher congealing accuracy, and a smaller SOF represents greater robustness. In addition, the efficiency of the algorithms is evaluated by the number of iterations to converge and the CPU time, which includes the times for both feature selection and congealing.

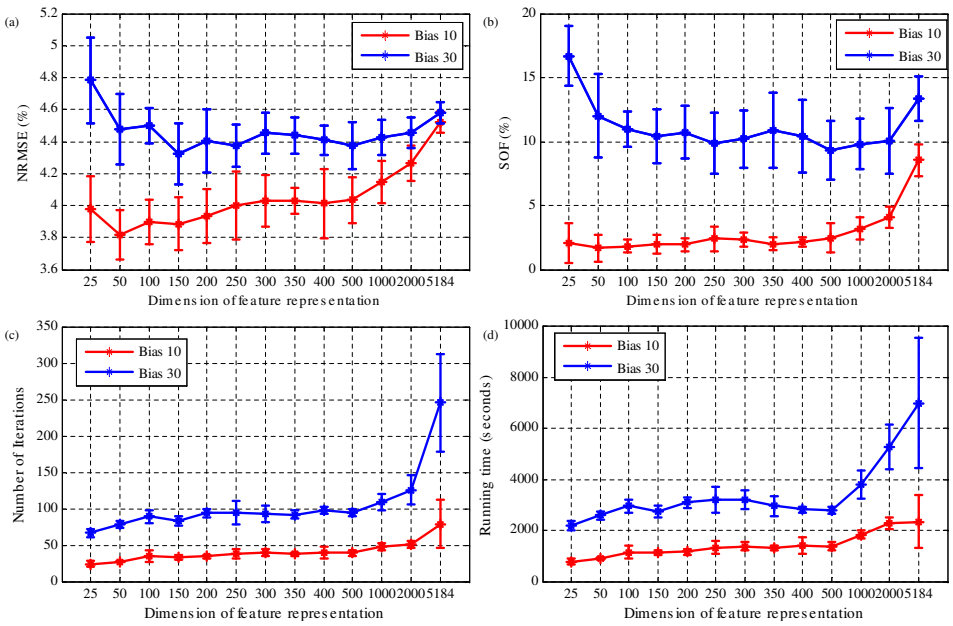


Figure 4. Congealing accuracy (a,b) and efficiency (c,d) with various feature dimension k. $k = 5184$ refers to the conventional congealing.

By setting $\eta_{max} = 10$, we first generate 5 random initializations for the 300-image ensemble. For each initialization, we specify various numbers of features (k) for the congealing algorithm to run. The same experiments are conducted for $\eta_{max} = 30$, which is more of an extreme case study because the commercial face detector by Pittsburgh Pattern Recognition [33] can achieve $\eta_{max} = 15$. Figure 4 shows the results where each dot and its variance are computed from 5 runs.

A number of observations can be made. For both cases of initialization, there is a large range of selected feature dimension (*e.g.*, $k \in [150, 500]$), from which the proposed algorithm achieves improved accuracy compared to the one without feature selection ($k = 5184$). This is a favorable property in that our algorithm is not sensitive to k. For both initializations, the new congealing always converges in less iterations and utilizes less CPU time, especially when k decreases. In the optimal case, when $\eta_{max} = 10$, our algorithm reduces the NRMSE from 4.5% to 3.8%, the SOF from 8.7% to 1.8%, and CPU time from $2,349$ to 912 seconds by merely using $\frac{50}{5184} = 0.96\%$ of the original feature. Comparing two cases of initialization, the improvement margin of accuracy by our algorithm in $\eta_{max} = 30$ is less than that of $\eta_{max} = 10$. This is partially due to the fact that the larger deviation at the initialization makes it challenging to converge by using a lower-dimensional feature representation. Hence, it might be wise to have the feature selection algorithm automatically nominate the optimal k at each congealing iteration, of interest for future work.

In addition to the quantitative evaluation, we also display the average warped image after congealing converges,

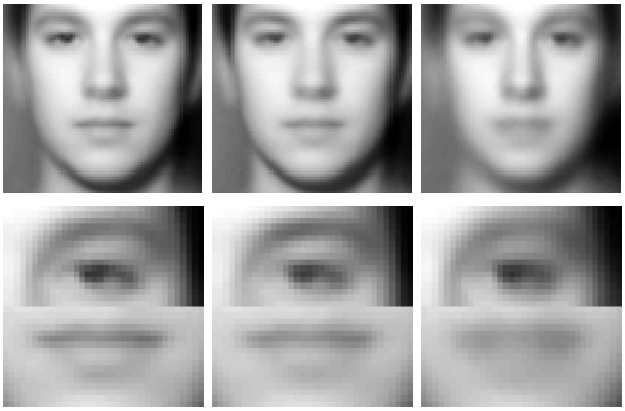


Figure 5. From left to right, average warped images at $\eta_{max} = 10$ when congealed with $k = 50$, 5184, and at the initialization.

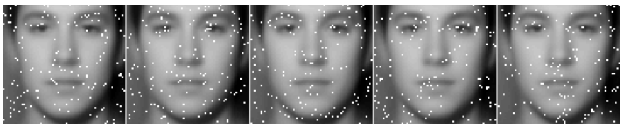


Figure 6. Selected feature locations at iteration #1,18,35,52,69.

which is expected to be sharp. From Fig. 5, we can see the improved sharpness when comparing $k = 50$ to $k = 5184$, especially in the eye and mouth regions.

Figure 6 plots the locations of the selected features at 5 iterations when $\eta_{max} = 10$ and $k = 50$. Notice that at different iterations, distinctive features are selected, many of which are co-located with facial features. For areas with relatively uniform appearance, such as cheek, fewer features

are chosen due to higher redundancy.

7. Conclusions

With the massive image data available for various object classes, image congealing is a key technology to automatically estimate the rigid or non-rigid deformation of the object instances. Armed with efficient unsupervised feature selection, the proposed congealing algorithm opens the potential of effectively performing congealing for a large image ensemble, despite the high dimensionality in the original feature representation. We show that with merely 3% of the original features, the proposed congealing can complete in less than 40% of the time, yet still improve the accuracy and robustness of congealing, when compared with conventional congealing without feature selection.

References

[1] C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2:1152–1174, 1974. 188

[2] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *IJCV*, 56(3):221–255, March 2004. 185, 187

[3] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem. *IEEE T-PAMI*, 26(10):1380–1384, October 2004. 186

[4] S. Balci, P. Golland, M. Shenton, and W. Wells. Free-form B-spline deformation model for groupwise registration. In *MICCAI*, pages 23–30, 2007. 186

[5] D. Blei and M. Jordan. Variational methods for the Dirichlet process. In *ICML*, 2004. 188

[6] K. I. Chang, K. W. Bowyer, and P. J. Flynn. An evaluation of multimodal 2D+3D face biometrics. *IEEE T-PAMI*, 27(4):619–624, 2005. 190

[7] T. Cootes, S. Marsland, C. Twining, K. Smith, and C. Taylor. Groupwise diffeomorphic non-rigid registration for automatic model building. In *ECCV*, volume 4, pages 316–327, 2004. 186

[8] T. Cootes, C. Twining, V. Petrovic, R. Schestowitz, and C. Taylor. Groupwise construction of appearance models using piece-wise affine deformations. In *BMVC*, volume 2, pages 879–888, 2005. 186

[9] M. Cox, S. Sridharan, S. Lucey, and J. Cohn. Least squares congealing for unsupervised alignment of images. In *CVPR*, 2008. 185, 186, 187

[10] M. Cox, S. Sridharan, S. Lucey, and J. Cohn. Least-squares congealing for large numbers of images. In *CVPR*, 2009. 186

[11] D. Cristinacce and T. Cootes. Facial motion analysis using clustered shortest path tree registration. In *Proc. of the 1st Intl. Workshop on Machine Learning for Vision-based Motion Analysis with ECCV*, 2008. 186

[12] N. Dalal and W. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893, 2005. 185

[13] M. Devaney and A. Ram. Efficient feature selection in conceptual clustering. In *ICML*, pages 92–97, 1997. 186

[14] J. G. Dy and C. E. Brodley. Feature subset selection and order identification for unsupervised learning. In *ICML*, pages 247–254, 2000. 186, 189

[15] C. Elkan. Using the triangle inequality to accelerate k-means. In *ICML*, pages 147–153, 2003. 188

[16] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999. 189

[17] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. of Machine Learning Research*, 3:1157–1182, March 2003. 186

[18] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *NIPS*, 2005. 186, 189

[19] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007. 185

[20] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *ICML*, pages 121–129, 1994. 186

[21] I. Kokkinos and A. Yuille. Unsupervised learning of object deformation models. In *ICCV*, 2007. 186

[22] M. H. Law, A. K. Jain, and M. A. T. Figueiredo. Feature selection in mixture-based clustering. In *NIPS*, pages 625–632, 2003. 186

[23] E. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE T-PAMI*, 28(2):236–250, 2006. 185, 186

[24] F. Lin and W. W. Cohen. Power iteration clustering. In *ICML*, 2010. 186, 188

[25] X. Liu, Y. Tong, and F. W. Wheeler. Simultaneous alignment and clustering for an image ensemble. In *ICCV*, 2009. 186

[26] X. Liu, Y. Tong, F. W. Wheeler, and P. H. Tu. Facial contour labeling via congealing. In *ECCV*, 2010. 185, 186

[27] M. Maila and J. Shi. A random walks view of spectral segmentation. In *AI and STATISTICS (AISTATS)*, 2001. 188

[28] E. Miller, N. Matsakis, and P. Viola. Learning from one example through shared densities on transforms. In *CVPR*, volume 1, pages 464–471, 2000. 185, 186

[29] P. Mitra, S. Member, C. A. Murthy, and S. K. Pal. Unsupervised feature selection using feature similarity. *IEEE T-PAMI*, 24:301–312, 2002. 186, 187, 189

[30] D. S. Modha and W. S. Spangler. Feature weighting in k-means clustering. *Machine Learning*, 52:217–237, September 2003. 186, 189

[31] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust batch alignment of images by sparse and low-rank decomposition. In *CVPR*, 2010. 186

[32] V. Roth and T. Lange. Feature selection in clustering problems. In *NIPS*, 2004. 186

[33] H. Schneiderman. Feature-centric evaluation for efficient cascaded object detection. In *CVPR*, volume 2, pages 29–36. IEEE, June 2004. 191

[34] J. Shi and J. Malik. Normalized cuts and image segmentation. In *CVPR*, pages 731–737, 1997. 188

[35] K. Sidorov, S. Richmond, and D. Marshall. An efficient stochastic approach to groupwise non-rigid image registration. In *CVPR*, 2009. 186

[36] M. Storer, M. Urschler, and H. Bischof. Intensity-based congealing for unsupervised joint image alignment. In *ICPR*, pages 569–576, 2010. 186

[37] Y. Tong, X. Liu, F. W. Wheeler, and P. Tu. Automatic facial landmark labeling with minimal supervision. In *CVPR*, 2009. 185, 186, 187, 189

[38] F. Torre and M. Nguyen. Parameterized kernel principal component analysis: Theory and applications to supervised and unsupervised image alignment. In *CVPR*, 2008. 186

[39] T. Vetter, M. J. Jones, and T. Poggio. A bootstrapping algorithm for learning linear models of object classes. In *CVPR*, pages 40–46, 1997. 186

[40] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, May 2004. 185

[41] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *J. of Machine Learning Research*, 5:1205–1224, December 2004. 186

[42] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *ICML*, pages 1151–1157, 2007. 186, 189