Efficient Face Alignment via Locality-constrained Representation for Robust Recognition

Yandong Wen[1*], Weiyang Liu[2*], Meng Yang[3], and Zhifeng Li[4]
[1]School of Electronic and Information Engineering, South China University of Technology
[2]School of Electronic and Computer Engineering, Peking University
[3]School of Computer Science and Software Engineering, Shenzhen University
[4]SIAT, Chinese Academy of Sciences

Abstract

Practical face recognition has been studied in the past decades, but still remains an open challenge. Current prevailing approaches have already achieved substantial breakthroughs in recognition accuracy. However, their performance usually drops dramatically if face samples are severely misaligned. To address this problem, we propose a highly efficient misalignment-robust locality-constrained representation (MRLR) algorithm for practical real-time face recognition. Specifically, the locality constraint that activates the most correlated atoms and suppresses the uncorrelated ones, is applied to construct the dictionary for face alignment. Then we simultaneously align the warped face and update the locality-constrained dictionary, eventually obtaining the final alignment. Moreover, we make use of the block structure to accelerate the derived analytical solution. Experimental results on public data sets show that MRLR significantly outperforms several state-of-the-art approaches in terms of efficiency and scalability with even better performance.

Introduction

Over the past years, face recognition has been and is still one of the most important and fundamental computer vision problem. Significant progresses have been made in face recognition, ranging from the family of sparse representation (Wright et al. 2009; Wagner et al. 2012; Zhang, Yang, and Feng 2011) to the application of deep convolutional neural network (CNN) (Sun, Wang, and Tang 2014a; Sun et al. 2014; Taigman et al. 2014). While achieving impressive recognition accuracy in controlled environments (some of them even surpass the human performance at certain tasks), most of them also show strong robustness to occlusion and illumination. However, these algorithms largely depend on well-aligned training and testing samples. Research (Shan et al. 2004) has demonstrated that even slight misalignment can globally transform the entire images, greatly reducing the recognition accuracy. Even the CNN that achieves the state-of-the-art performance nowadays needs to align the training and testing faces to the same position, since misaligned query faces can greatly degrade its performance (Schroff, Kalenichenko, and Philbin 2015).

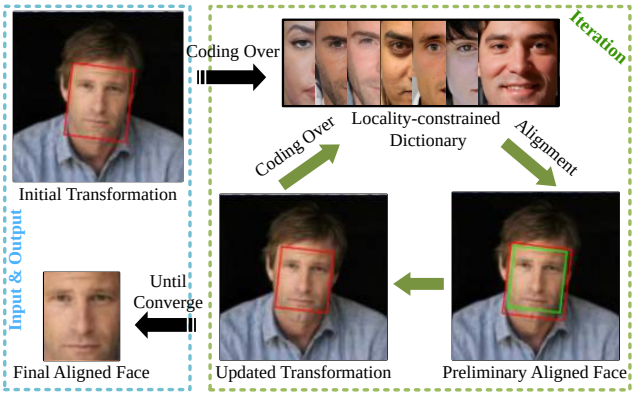

Figure 1. An Illustration of the MRLR. The red bounding box denotes the input estimate and the green one is the output estimate. The whole algorithm works in an iterating style.

Thus current face recognition techniques can benefit from an efficient and well-performing face alignment algorithm.

In this paper, we only consider the face alignment methods based on subspace learning and sparse representation (Huang, Huang, and Metaxas 2008; Yang, Zhang, and Zhang 2012; Wagner et al. 2012). Although there are other types of face image registration methods that can handle larger face variation in expression and pose, e.g. active appearance models (Cootes, Edwards, and Taylor 2001), active shape models (Cootes et al. 1995) and unsupervised joint alignment (Huang, Jain, and Learned-Miller 2007), their complexity is usually too high for efficient alignment while ours is far more efficient and suitable for real-time situations. Besides, the facial landmarks based methods only focus on accurately detect the facial key points, while ours aligns the face based on the whole training samples and focus on benefiting the subsequent recognition.

Related Work

(Wright et al. 2009) reported the sparse representation based classification (SRC), which seeks to represent an aligned testing image by the linear sparse combination of training images. The basic assumption for SRC is that all the training and testing samples need to be well aligned, so SRC performs poorly with misaligned faces. To overcome such shortcoming, (Huang, Huang, and Metaxas 2008) proposed

* indicates equal contributions.

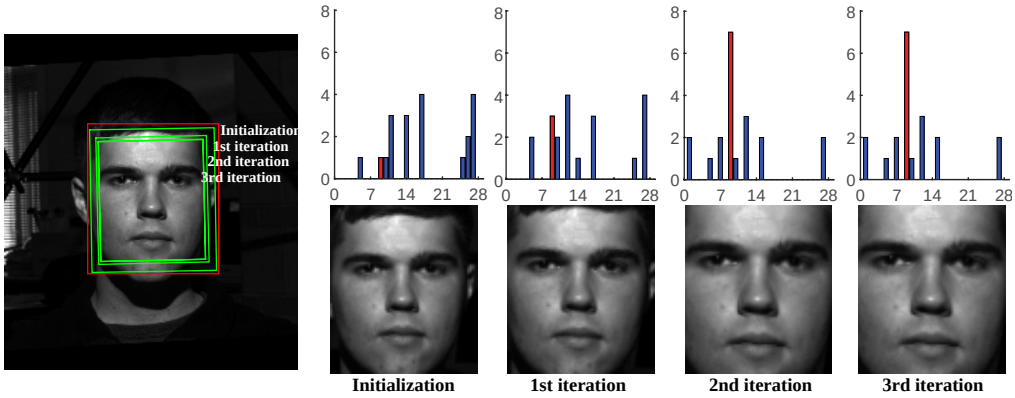

Figure 2. An Illustration of the MRLR iteration procedure. The left image is the input uncropped face. We use the Viola-Jones detector to generate an initialized estimate, and obtain the locality-constrained dictionary (LCD) for it, while the LCD is to compute the new transformation. Then we iteratively update the face transformation and the corresponding LCD until convergence. The bar plot denotes the label distribution of the LCD. It shows that the LCD contains increasingly more training samples from the same class as the testing face after each iteration.

the transform-invariant sparse representation (TSR). They add deformations in training set, simultaneously recovering the image transformation and representation coefficients. However, TSR aligns testing image to global dictionary and thus easily gets trapped in local minima. To avoid that, robust alignment by sparse representation (RASR) (Wagner et al. 2012) aligns the testing image to training samples of each subject, then warps training set and testing image to a unified transformation for recognition. The exhaustive subject by subject search effectively finds the global optima, but it is extremely time-consuming, especially when the subject number is large. Therefore RASR detrimental to efficiency and scalability. (Yang, Zhang, and Zhang 2012) proposed the efficient misalignment-robust representation (MRR) for face recognition. With the carefully controlled training set, they perform the singular value decomposition (SVD) and use principal components to approximate the global dictionary, significantly enhancing its real-time ability. However, SVD operation therein is still time and memory consuming, preventing MRR from being applied in large-scale datasets. Using the principle components of dictionary instead of the original one inevitably reduces alignment accuracy.

Motivations and Contributions

In summary, current prevailing sparse representation based face alignment methods contain several major shortcomings:

- Time-consuming: (Wagner et al. 2012; Zhuang et al. 2013) need to align the face in an exhaustive manner using sub-dictionaries that are constructed by every individual. Suppose the dataset contains more than a thousand individuals, these algorithms will work extremely slow.
- Easy to introduce background noise: (Wagner et al. 2012; Zhuang et al. 2013) align the training set to the testing sample (e.g. (Wagner et al. 2012; Zhuang et al. 2013)), which may introduce background noise if the testing sample is largely off-centered and break the low-dimensional linear illumination model (Basri and Jacobs 2003).
- Unsatisfactory subspace: (Huang, Huang, and Metaxas 2008; Yang, Zhang, and Zhang 2012) use the global dictionary to perform the alignment. The global dictionary contains various uncorrelated face samples and produces a unsatisfactory subspace for alignment.
- Unable to benefit from outside data.

In order to address the above problems, we propose a misalignment-robust locality-constrained representation (MRLR) for robust face recognition. Fig. 1 briefly illustrates the MRLR. Inspired by the locality-constrained linear coding (Wang et al. 2010), the locality is introduced to the dictionary construction for alignment. Specifically, we combine a locality adaptor to the l_2 regularized penalties for $\boldsymbol{x}$. Because we also use l_2 norm to constrain $\boldsymbol{e}$, an efficient analytical solution can be derived. While updating the face transformation, we simultaneously update the locality-constrained dictionary, as shown in Fig. 2. Our contributions are summarized as follows.

- The proposed locality-constrained representation avoid the exhaustive search in every subject of the training set, greatly reducing the computational time and making the alignment scalable to large datasets. To the best of our knowledge, this is the first time that locality has been introduced to improve the performance of face alignment.
- MRLR uses the locality adaptor and the l_2-norm to penalize the the representation term and the error term. We derive an analytical solution for the optimization. Moreover, we can accelerate the analytical solution by making use of the block structure of the deformable dictionary. Thus the inverse of a large-size matrix can be further avoided, making our model even more scalable and efficient.
- MRLR simultaneously optimize the transformation and update the corresponding locality-constrained dictionary, which largely avoids the unsatisfactory local minima.
- MRLR can take advantage of outside data to better construct the locality-constrained dictionary. Outside data can be effectively used to benefit the alignment performance.

The Proposed Method

The MRLR Model

We arrange the given n_i training samples from the ith class as columns of a dictionary $\boldsymbol{D}_i = [\boldsymbol{d}_{i,1}, \boldsymbol{d}_{i,2}, \cdots, \boldsymbol{d}_{i,n_i}] \in$

(a) L1, Running Time: 6.850s

(b) L2, Running Time: 2.341s

(b) Uncorresponding Sub-Dictionary, Running Time: 1.211s

(d) Corresponding Sub-Dictionary, Running Time: 1.131s

(e) Facial Landmarks based Methods, Running Time: 1.596s

(f) MRLR, Running Time: 0.636s

Figure 3. A face alignment example using different constraints, key point based method and MRLR. The red box is the initial estimate and the green box denotes the final alignment. After performing face alignment, we use the SRC to perform the face recognition. The stem diagram on the right shows the corresponding sparse representation coefficients of the aligned face after performing SRC. We can see only (d) and (f) show discriminative representation results. Note that, the corresponding sub-dictionary in (d) means we only use the sub-dictionary that is constituted by faces whose label is the same as the testing face. The uncorresponding sub-dictionary in (c) is constituted by faces that do not belong to the testing face.

$\mathbb{R}^{m\times n_i}$ where $\boldsymbol{d}_{i,j} \in \mathbb{R}^m$ denotes the jth vectorized training sample of the ith class. Combining all the dictionary from each subjects, we can obtain a global dictionary $\boldsymbol{D} = [\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_k] \in \mathbb{R}^{m\times n}$ where $n = \sum_{i=1}^{k} n_i$ and k is the number of subjects. Suppose that the query face $\boldsymbol{y}$ belongs to the ith subject, ideally it can be approximately represented by $\boldsymbol{D}_i\boldsymbol{x}$ in which $\boldsymbol{x}$ is the representation coefficients. However, due to the misalignment problem, such linear subspace representation may be invalid. Therefore we introduce a transformation that models the warping to the original face. Instead of observing the $\boldsymbol{y}$, we observe the warped face $\boldsymbol{y}_w = \boldsymbol{y}\circ\tau^{-1}$ where $\circ$ denotes a nonlinear operator and τ belongs to a finite-dimensional group of transformations acting on the image domain (e.g. similarity transformation). The linear subspace representation $\boldsymbol{x}$ of the warped face can not reveal the true identity. Naively applying recognition algorithm is inappropriate. On the other hand, the potential subspace corresponding to $\boldsymbol{y}$ is also unknown, so it is difficult to align it. Fortunately, by leveraging the high similarity of face, we can construct a suboptimal local dictionary for alignment, and update the local dictionary according to the latest transformation. After several iterations, it eventually converges to the accurate transformation. After the true deformation τ^{-1} is found, then we can apply its inverse τ to the testing face and obtain the aligned face $\boldsymbol{y}_w \circ \tau$.

The global dictionary with l_1/l_2 constraint usually recovers the unsatisfactory transformation (see Fig. 3), because it is prone to local minima under the interference of atoms from the other subjects. Inspired by (Wang et al. 2010), we introduce the locality constraints to the dictionary. The reason lies in two folds. First, locality-constrained dictionary only uses the most similar atoms to the query, effectively avoiding unsatisfactory local minima caused by dissimilar atoms, as shown in Fig. 3. Second, using locality-constrained dictionary requires no exhaustive search in every subject and leads to highly efficient solving algorithm. The model of MRLR is formulated as

$$\min_{\boldsymbol{x},\boldsymbol{e}} \|\boldsymbol{c}\odot\boldsymbol{x}\|_2^2 + \|\boldsymbol{e}\|_2^2 \quad \text{s.t. } \boldsymbol{y}_w\circ\tau = \boldsymbol{D}\boldsymbol{x} + \boldsymbol{e} \tag{1}$$

where $\odot$ denotes the element-wise multiplication between two vectors, and $\boldsymbol{c} \in \mathbb{R}^n$ is the locality adaptor that attaches different penalties to the coefficients $\boldsymbol{x}$. The locality adaptor activates the most correlated atoms for the testing sample, while suppressing the uncorrelated ones. Unfortunately, the model in Eq. (1) is difficult to solve due to the non-linearity. A small deformation in the transform τ can be linearized as $\boldsymbol{y}_w \circ (\tau + \Delta\tau) = \boldsymbol{y}_w \circ \tau + J\Delta\tau$ where $J = \frac{\partial}{\partial r}\boldsymbol{y}_w \circ \tau$ is the Jacobian of $\boldsymbol{y}\circ\tau$ with respect to τ and $\Delta\tau$ is the step in τ. If an initial τ is given, we can repeatedly search for an optimal $\Delta\tau$ to update τ and J. A final transformation τ can be obtained to align the warped image.

The efficiency of the MRLR model lies in two folds. First, we enforce the l_2 norm constraints on both $\boldsymbol{c}\odot\boldsymbol{x}$ and $\boldsymbol{e}$, and derive an analytical solution for MRLR, which is much faster than solving l_1 norm minimization. In fact, the performance of the l_2 constraints are similar to the l_1 constraints in the case without occlusion (Zhang, Yang, and Feng 2011). Second, we take advantage of the block structure of matrices to design a highly efficient algorithm, which obtains exactly the same solution in shorter time. The MRLR algorithm is

summarized as follows.

Algorithm 1 The MRLR algorithm for Face Alignment

Input: The dictionary of training samples $\boldsymbol{D}$, the warped testing image $\boldsymbol{y}_w$, the initial transformation τ (it can be obtained by any off-the-shelf face detector, e.g. Viola-Jones detector), a constant σ.

Output: The aligned face $\boldsymbol{y}$

1: **while** not converge or reach maximal iteration **do**
2: Compute the locality adaptor: $\boldsymbol{c} \leftarrow \exp(\frac{\boldsymbol{D}^T\boldsymbol{y}}{\sigma})$, for all i, $\boldsymbol{c}_i \leftarrow max(\boldsymbol{c}) - \boldsymbol{c}_i$.
3: $j \leftarrow 1$.
4: **while** not converge or reach maximal iteration **do**
5: $\hat{\boldsymbol{y}}_w(\tau_{j-1}) \leftarrow \frac{\boldsymbol{y}_w \circ \tau_{j-1}}{\|\boldsymbol{y}_w \circ \tau_{j-1}\|_2}$, $J \leftarrow \frac{\partial}{\partial \tau_{j-1}} \hat{\boldsymbol{y}}_w(\tau_{j-1})|_{\tau_{j-1}}$.
6: $\Delta\tau = \arg\min_{\Delta\tau, \boldsymbol{x}, \boldsymbol{e}} \|\boldsymbol{c} \odot \boldsymbol{x}\|_2^2 + \|\boldsymbol{e}\|_2^2$ s.t. $\hat{\boldsymbol{y}}_w(\tau_j) + J\Delta\tau = \boldsymbol{D}\boldsymbol{x} + \boldsymbol{e}$.
7: $\tau_j \leftarrow \tau_{j-1} + \Delta\tau$.
8: $j \leftarrow j + 1$.
9: **end while**
10: $\tau \leftarrow \tau_j$, $\tau_0 \leftarrow \tau_j$.
11: **end while**
12: Output the final aligned face $\boldsymbol{y} = \boldsymbol{y}_w \circ \boldsymbol{e}$.

Efficient Solving Algorithm

This section presents a highly efficient solution for the MRLR algorithm. By analyzing Algorithm 1, we find the optimization in step 6 dominates the overall computational time. Although it has an analytical solution, it contains the inversion operation of a large-size matrix. We aim to take advantage of the block structure of the matrix to decompose the inversion. We first reformulate the optimization in Step 6 as

$$\begin{aligned} \Delta\boldsymbol{\tau} &= \min_{\Delta\tau} \|\boldsymbol{C}\boldsymbol{x}\|_2^2 + \|\boldsymbol{e}\|_2^2 \\ s.t.\ \ & \hat{\boldsymbol{y}}_w + J\Delta\tau = \boldsymbol{D}\boldsymbol{x} + \boldsymbol{e} \end{aligned} \tag{2}$$

where $\boldsymbol{C}$ is a diagonal matrix with the diagonal elements being the locality adaptor vector $\boldsymbol{c}$. We can further substitute $\boldsymbol{e} = \hat{\boldsymbol{y}}_w - [\boldsymbol{D}, -J] \begin{bmatrix} \boldsymbol{x} \\ \Delta\tau \end{bmatrix}$ into Eq. (2), we have

$$\begin{aligned} \Delta\tau &= \arg\min_{\boldsymbol{x}, \Delta\tau} \|\boldsymbol{C}\boldsymbol{x}\|_2^2 + \left\| \hat{\boldsymbol{y}}_w - [\boldsymbol{D}, -J] \begin{bmatrix} \boldsymbol{x} \\ \Delta\tau \end{bmatrix} \right\|_2^2 \\ &= \arg\min_{\boldsymbol{x}, \Delta\tau} \left\| \begin{bmatrix} \hat{\boldsymbol{y}}_w \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{D} & -J \\ \boldsymbol{C} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \Delta\tau \end{bmatrix} \right\|_2^2 \\ &= \arg\min_{\boldsymbol{z}} \|\boldsymbol{u} - R\boldsymbol{z}\|_2^2 \end{aligned} \tag{3}$$

where $\boldsymbol{u}$, $\boldsymbol{R}$ and $\boldsymbol{z}$ denote $\begin{bmatrix} \hat{\boldsymbol{y}}_w \\ \boldsymbol{0} \end{bmatrix}$, $\begin{bmatrix} \boldsymbol{D} & -J \\ \boldsymbol{C} & \boldsymbol{0} \end{bmatrix}$ and $\begin{bmatrix} \boldsymbol{x} \\ \Delta\tau \end{bmatrix}$ respectively. It becomes a least square problem whose analytical solution is $\boldsymbol{z} = (\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T\boldsymbol{u}$. As one can see, the computational complexity is still high due to the large size of $\boldsymbol{R}$. Actually, the efficiency and the scalability can be greatly boosted if we make good use of the block structure of the matrix $\boldsymbol{R}$.

Using the block matrix inversion, we can rewrite the analytical solution as

$$\begin{aligned} \boldsymbol{z} &= (\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T\boldsymbol{u} \\ &= \left(\begin{bmatrix} \boldsymbol{D}^T & \boldsymbol{C}^T \\ -J^T & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{D} & -J \\ \boldsymbol{C} & \boldsymbol{0} \end{bmatrix} \right)^{-1} \begin{bmatrix} \boldsymbol{D}^T & \boldsymbol{C} \\ -J^T & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{y}}_w \\ \boldsymbol{0} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{D}^T\boldsymbol{D} + \boldsymbol{C}^T\boldsymbol{C} & -\boldsymbol{D}^TJ \\ -J^T\boldsymbol{D} & J^TJ \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{D}^T \\ -J^T \end{bmatrix} \hat{\boldsymbol{y}}_w \\ &= \begin{bmatrix} \boldsymbol{Z}_1^{-1} & (\boldsymbol{D}^T\boldsymbol{D} + \boldsymbol{C}^T\boldsymbol{C})^{-1} \\ & \times (\boldsymbol{D}^TJ)\boldsymbol{Z}_2^{-1} \\ \boldsymbol{Z}_2^{-1}(J^T\boldsymbol{D})\times & \\ (\boldsymbol{D}^T\boldsymbol{D} + \boldsymbol{C}^T\boldsymbol{C})^{-1} & \boldsymbol{Z}_2^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{D}^T \\ -J^T \end{bmatrix} \hat{\boldsymbol{y}}_w \end{aligned} \tag{4}$$

We denote $\boldsymbol{D}^T\boldsymbol{D} + \boldsymbol{C}^T\boldsymbol{C}$, $\boldsymbol{D}^TJ$ and J^TJ as $\boldsymbol{T}_1$, $\boldsymbol{T}_2$ and $\boldsymbol{T}_3$ respectively. In particular, $\boldsymbol{T}_1$ and $\boldsymbol{T}_1^{-1}$ can be pre-calculated before the inner iteration from Step 4 to Step 9. The other variables $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ can be represented as $\boldsymbol{Z}_1^{-1} = (\boldsymbol{T}_1 - \boldsymbol{T}_2\boldsymbol{T}_3^{-1}\boldsymbol{T}_2^T)^{-1}$ and $\boldsymbol{Z}_2^{-1} = (\boldsymbol{T}_3 - \boldsymbol{T}_2^T\boldsymbol{T}_1^{-1}\boldsymbol{T}_2)^{-1}$. Eq. (4) can be represented as

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \Delta\tau \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_1^{-1}(\boldsymbol{D}^T\hat{\boldsymbol{y}}_w) - \boldsymbol{T}_1^{-1}\boldsymbol{T}_2\boldsymbol{Z}_2^{-1}(J^T\hat{\boldsymbol{y}}_w) \\ \boldsymbol{Z}_2^{-1}\boldsymbol{T}_2^T\boldsymbol{T}_1^{-1}(\boldsymbol{D}^T\hat{\boldsymbol{y}}_w) - \boldsymbol{Z}_2^{-1}(J^T\hat{\boldsymbol{y}}_w) \end{bmatrix} \tag{5}$$

Note that the purpose of the face alignment is to search a deformation step $\Delta\tau$, so computing $\boldsymbol{x}$ is unnecessary. Without computing $\boldsymbol{x}$, we can save greatly reduce the computation. Moreover, as mentioned in (Wang et al. 2010), since $\boldsymbol{c}$ usually imposes weak constraint on only a few atoms, suppressing most of the atoms. We can simply keep the smallest $s, (s \ll n)$ entries in $\boldsymbol{c}$ and force other entries to be positive infinity. This strategy further accelerates the coding, as we present in complexity analysis and experiments (This strategy is termed as MRLR2, while the former proposed one is termed as MRLR1). Detailed complexity analysis refers to the supplementary material

Experiments

We conduct experiments on the face database (Extended Yale B (Georghiades, Belhumeur, and Kriegman 2001) and CAS-PEAL (Gao et al. 2008)) with controlled laboratory conditions to comprehensively evaluate MRLR in terms of region of attraction, recognition rate, running time and scalability. Then practical face recognition performance are evaluated by the Labeled Faces in the Wild (LFW) dataset (Huang et al. 2007). The experimental results show that MRLR achieves competitive performance with much less running time and scales better in large datasets. Moreover, MRLR is able to make use of outside data to improve alignment, benefiting the subsequent recognition in the scenario where only one sample each person is available.

Implementation details

In MRLR2 and MRR, the length (the number of atoms) of dictionary for alignment is fixed to 20 for fair comparison. We basically follow the same setting in (Wagner et al. 2012), 10 classes after first stage are remained in MRR and RASR, and one project matrix of 500 rows is used in TSR. The illumination dictionary in (Zhuang et al. 2013) follows its original setup, and the amount of illumination atoms is 30 in

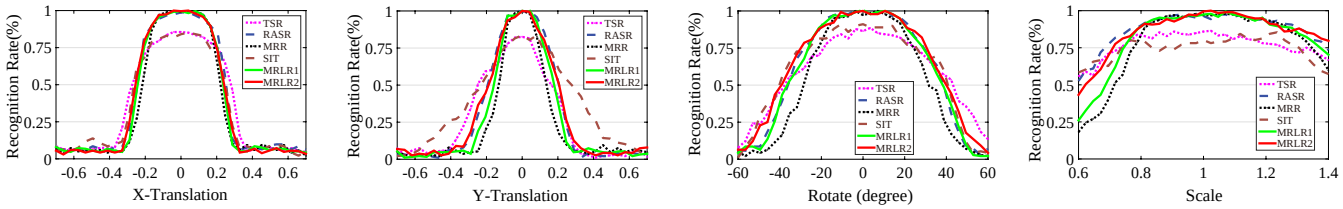

Figure 4. The region of attraction. (The amount of translation is given as a fraction of the distance between eyes) (a) Translation in the Y-direction only. (b) Translation in the X-direction only. (c) In-plane rotation only. (d) Scale variation only.

all experiments. The maximum iteration of outer and inner loop for MRLR in these methods are consistently set to 3 and 30.. The l_1-minimization algorithm uses the Augmented Lagrange Multiplier (Yang et al. 2010).

The region of attraction

The region of attraction evaluates the robustness against 2D deformations. We compare MRLR with TSR (Huang, Huang, and Metaxas 2008), RASR (Wagner et al. 2012), MRR (Yang, Zhang, and Zhang 2012) and SIT (Zhuang et al. 2013) on Extended Yale B database, which includes 2414 images of 38 subjects. We use the uncropped images of 28 subjects in experiments. 32 training images per category are randomly selected, and the rest are used for testing. All the training images are resized to 80×70. We get access to the ground truth of eyes and add perturbation to them. Then we calculate the corresponding recognition accuracy under various initial transformations. One can see that MRLR performs well and stably within a certain range of misalignment, e.g. 20 percent translation in x direction (14-16 pixels), 20 degree rotation or 30 percent scale variation. It significantly enhances the robustness of practical recognition, because the average misalignment of a face detector safely falls within 10 percent translation and 8 percent scale variation. TSR performs relatively poor even in small perturbations. It is mainly because aligning testing image to the entire training set is more prone to local minima, resulting in inaccurate alignment. Using single sample per class, SIT achieves similar performance with TSR due to the limited representation ability. Compared to MRLR1, MRLR2 and RASR, MRR performs slightly worse on robustness to deformation. MRLR1 and MRLR2 perform almost the same as RASR, demonstrating that locality-constrained representation effectively avoids local minima.

Face recognition in controlled environments

We conduct the recognition experiments on both Extended Yale B and CAS-PEAL datasets. For Extended Yale B, we adopt the same settings in the previous section. For CAS-PEAL, 20 subjects were chosen, each of them including more than 32 images. We randomly selected 20 images per subject and resized them to 80×70 for training, then test on the remaining 12 images. Because SIT (Zhuang et al. 2013) trains on single sample per category, we also reduce the training set in MRLR to single sample per category for fair comparison (termed as MRLR-SS). The initial τ_0 are automatically given by Viola-Jones detector (Viola and Jones 2001). Table 1 gives the recognition rates and average running time.

As we discuss above, unsatisfactory local minima in global dictionary leads to poor performance (81.61%) in TSR. With less amount of subject (from 32 to 20), local minima is alleviated and TSR is able to perform better. RASR performs very well in both Extended Yale B dataset (92.42%) and CAS-PEAL dataset (89.92%). However, such subject by subject search is time-consuming, it averagely costs 9.76 and 5.45 seconds on each testing image when the amount of subjects are 28 and 20, respectively. On the other hand, the recognition rate of MRLR2 is 92.68% and 90.43%, slightly better than RASR. Most importantly, it takes only 0.18 and 0.15 seconds to deal with a testing image, roughly 4, 55 and 41 times faster than MRR, RASR and TSR respectively. With single sample each subject, SIT achieves 84.53% and 86.76% recognition rate in two datasets respectively, better than the single sample version of MRLR. Because the dictionary for alignment consists of illumination dictionary (outside samples) and single training sample per class, it shares the same scale with RASR, resulting in similar running time.

Scalability

We vary the number of subject from 10 to 100 and resize the images from 40×35 to 160×140, to evaluate the scalability of our algorithm. Table 2 and Table 3 show the experimental results. One can observe that TSR, RASR and SIT cost too much time, far from being applicable in real-time systems. The running time of TSR remains relatively stable as the dimension increases, but rises linearly with more subjects. MRR maintains excellent real-time capability with the growth of the subject number. However, its running time rises dramatically when the resolution of image increasing. Unlike the abovementioned approaches, MRLR1 and MRLR2 are not very sensitive to the dimension or number of subjects, preserving competitive performance. MRLR2

Table 1. The recognition accuracy and running time on Extended Yale B and CAS-PEAL datasets.

Method	Extended Yale B		CAS-PEAL	
	Recognition Rate	Running Time	Recognition Rate	Running Time
TSR	81.61	7.396	86.96	4.2695
RASR	92.42	9.7587	89.92	5.4466
MRR	90.95	0.7773	90.00	0.5684
SIT	84.53	9.9823	86.76	6.0329
MRLR-SS	77.12	**0.1566**	81.31	**0.1384**
MRLR1	92.31	0.6207	89.76	0.3307
MRLR2	**92.53**	0.1783	**90.43**	0.1462

costs the least running time and the lowest increasing rate as we enlarge the dimension or number of subjects, showing the best scalability among state-of-the-art approaches.

Table 2. Running time (s) under different dimensions (image size).

Method	40×35	64×56	80×70	120×105	160×140
TSR	3.645	3.861	4.270	4.672	5.468
RASR	3.499	4.452	6.110	10.324	17.111
MRR	0.133	0.342	0.593	2.259	5.997
SIT	3.564	4.637	6.565	11.035	19.215
MRLR1	0.085	0.195	0.331	0.569	0.940
MRLR2	**0.066**	**0.118**	**0.146**	**0.303**	**0.505**

Table 3. Running time (s) under different amount of classes.

Method	10	20	40	70	100
TSR	2.1533	3.2825	5.5280	8.4034	11.5327
RASR	2.7377	4.6596	8.8647	15.4644	22.1281
MRR	0.5776	0.5928	0.6082	0.6394	0.6994
SIT	2.86	5.1996	9.9817	17.6875	27.1734
MRLR1	0.1977	0.2819	0.513	0.8552	1.4096
MRLR2	**0.1318**	**0.1373**	**0.1559**	**0.197**	**0.2616**

Face recognition and verification in the wild

In this section, we test MRLR in practical scenario. LFW dataset contains 13,233 images of 5,749 people, while 4,069 people have only one image. This dataset is very challenging since it is collected in the uncontrolled wild scenario, including blur, various illumination, crossing age, occlusion or misalignment. We present two experiments, face recognition and face verification on LFW database to evaluate the performance of MRLR. In recognition testing, we choose 20 persons with more than 20 images, forming a subset with 1534 samples. We randomly select 20 samples each subject as training set, and test on the rest. The experimental results are shown in Table 3. It is worth noticing that there are many testing images including 3D deformation. Although aligning a 3D warped image to frontal face is beyond the scope of our approach, we do not manually exclude these images, for the purpose of evaluating the performance of our method in practical scenario. It is clear that MRLR performs best among these misalignment-robust recognition algorithms, outperforming RASR for 1.55% in recognition rate. Furthermore, the single sample version of MRLR beats SIT with a significant margin. It is mainly because the illumination dictionary is not informative enough to represent such sophisticated intra-class variation in each subject.

To address the problem of insufficient training images when aligning, we propose to use outside data to improve alignment. Making fully use of the similarity of face, the outside data that belong to neither the training subjects nor the testing subjects, also enhances the accuracy of alignment. With the outside data, MRLR performs better even in the scenario where there is only one sample per subject.

Face verification represents another task. Given two face images, the goal is to decide whether the two people pictured belong to the same individual. Many breakthroughs have been achieve by Convolutional Neural Network (CNN) (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014). However, the training data also need to be loosely or accurately aligned, so that the verification accuracy can be further boosted. In (Sun, Wang, and Tang 2014a; Sun et al. 2014; Sun, Wang, and Tang 2014b), similarity transformation is used to align training and testing images according the facial landmarks. (Taigman et al. 2014; Schroff, Kalenichenko, and Philbin 2015) also state that accurate 3D aligning do help the subsequent face verification. In this experiment, we train a simple neural network consisting of 7 convolution layers and 2 fully connected layers, jointly supervised by softmax loss and contrastive loss. The specific deep model description is given in the supplementary material. The deep model is trained on roughly 600 thousand outside samples (These samples and LFW do not share the same individuals.) and test on LFW, following the standard unrestricted protocol. The feature of each image are taken from the output of the first fully connected layer, and their Euclidean distance are calculated for binary classification. The pairs whose distance exceeds the threshold are regarded as negative. We compare landmarks based (IntraFace (Asthana et al. 2014)) method and MRLR by aligning the testing image, and carry out ten-fold cross validation testing on 6000 pairs. The results are reported in Table 5.

Table 4. Recognition accuracy (%) on LFW dataset.

Method	Accuracy	Method	Accuracy
TSR	73.63	**MRLR-SS**	72.47
RASR	81.43	**MRLR-SS** with outside data	80.42
MRR	78.84	**MRLR1**	**82.98**
SIT	55.91	**MRLR2**	81.75

Table 5. Verification accuracy (%) on LFW dataset.

Method	No. of points	Distance	Accuracy (%)
Intraface	5	L2	98.05±0.64
Intraface	5	PCA+L2	98.00±0.68
Intraface	12	L2	98.09±0.60
Intraface	12	PCA+L2	98.15±0.49
Intraface	49	L2	98.22±0.61
Intraface	49	PCA+L2	98.32±0.47
MRLR2	N/A(Image Set)	L2	**98.68±0.51**
MRLR2	N/A(Image Set)	PCA+L2	**98.77±0.45**

Concluding Remarks

In this paper, we propose an efficient misalignment-robust locality representation algorithm, MRLR, for face alignment. The locality constraint therein avoids the interference of the uncorrelated atoms and the exhaustive search in every subject, greatly reducing running time while still preserving accurate alignment. Moreover, motivated by the block structure of dictionary, we propose an efficient solving algorithm to speed up the alignment. Besides, MRLR is easily extended to one-shot face alignment and can benefit from outside data. Computational complexity analysis and extensive experiments show that MRLR considerably reduce the running time with even better performance.

References

[Asthana et al. 2014] Asthana, A.; Zafeiriou, S.; Cheng, S.; and Pantic, M. 2014. Incremental face alignment in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 1859–1866. IEEE.

[Basri and Jacobs 2003] Basri, R., and Jacobs, D. W. 2003. Lambertian reflectance and linear subspaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 25(2):218–233.

[Cootes et al. 1995] Cootes, T. F.; Taylor, C. J.; Cooper, D. H.; and Graham, J. 1995. Active shape models-their training and application. *Computer vision and image understanding* 61(1):38–59.

[Cootes, Edwards, and Taylor 2001] Cootes, T. F.; Edwards, G. J.; and Taylor, C. J. 2001. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (6):681–685.

[Gao et al. 2008] Gao, W.; Cao, B.; Shan, S.; Chen, X.; Zhou, D.; Zhang, X.; and Zhao, D. 2008. The cas-peal large-scale chinese face database and baseline evaluations. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 38(1):149–161.

[Georghiades, Belhumeur, and Kriegman 2001] Georghiades, A. S.; Belhumeur, P. N.; and Kriegman, D. J. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23(6):643–660.

[Huang et al. 2007] Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst.

[Huang, Huang, and Metaxas 2008] Huang, J.; Huang, X.; and Metaxas, D. 2008. Simultaneous image transformation and sparse representation recovery. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

[Huang, Jain, and Learned-Miller 2007] Huang, G. B.; Jain, V.; and Learned-Miller, E. 2007. Unsupervised joint alignment of complex images. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 1–8. IEEE.

[Krizhevsky, Sutskever, and Hinton 2012] Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

[Schroff, Kalenichenko, and Philbin 2015] Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. *arXiv preprint arXiv:1503.03832*.

[Shan et al. 2004] Shan, S.; Chang, Y.; Gao, W.; Cao, B.; and Yang, P. 2004. Curse of mis-alignment in face recognition: problem and a novel mis-alignment learning solution. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, 314–320. IEEE.

[Simonyan and Zisserman 2014] Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[Sun et al. 2014] Sun, Y.; Chen, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, 1988–1996.

[Sun, Wang, and Tang 2014a] Sun, Y.; Wang, X.; and Tang, X. 2014a. Deep learning face representation from predicting 10,000 classes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 1891–1898. IEEE.

[Sun, Wang, and Tang 2014b] Sun, Y.; Wang, X.; and Tang, X. 2014b. Deeply learned face representations are sparse, selective, and robust. *arXiv preprint arXiv:1412.1265*.

[Taigman et al. 2014] Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 1701–1708. IEEE.

[Viola and Jones 2001] Viola, P., and Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, I–511. IEEE.

[Wagner et al. 2012] Wagner, A.; Wright, J.; Ganesh, A.; Zhou, Z.; Mobahi, H.; and Ma, Y. 2012. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34(2):372–386.

[Wang et al. 2010] Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; and Gong, Y. 2010. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 3360–3367. IEEE.

[Wright et al. 2009] Wright, J.; Yang, A. Y.; Ganesh, A.; Sastry, S. S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31(2):210–227.

[Yang et al. 2010] Yang, A. Y.; Sastry, S. S.; Ganesh, A.; and Ma, Y. 2010. Fast l1-minimization algorithms and an application in robust face recognition: A review. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, 1849–1852. IEEE.

[Yang, Zhang, and Zhang 2012] Yang, M.; Zhang, L.; and Zhang, D. 2012. Efficient misalignment-robust representation for real-time face recognition. In *Computer Vision–ECCV 2012*. Springer. 850–863.

[Zhang, Yang, and Feng 2011] Zhang, L.; Yang, M.; and Feng, X. 2011. Sparse representation or collaborative representation: Which helps face recognition? In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 471–478. IEEE.

[Zhuang et al. 2013] Zhuang, L.; Yang, A. Y.; Zhou, Z.; Sastry, S. S.; and Ma, Y. 2013. Single-sample face recognition with image corruption and misalignment via sparse illumi-

nation transfer. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 3546–3553. IEEE.