Coupling Alignments with Recognition for Still-to-Video Face Recognition

Zhiwu Huang[1,2], Xiaowei Zhao[1,2], Shiguang Shan[1], Ruiping Wang[1], Xilin Chen[1]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
{zhiwu.huang, xiaowei.zhao, shiguang.shan, ruiping.wang, xilin.chen}@vipl.ict.ac.cn

Abstract

The Still-to-Video (S2V) face recognition systems typically need to match faces in low-quality videos captured under unconstrained conditions against high quality still face images, which is very challenging because of noise, image blur, low face resolutions, varying head pose, complex lighting, and alignment difficulty. To address the problem, one solution is to select the frames of 'best quality' from videos (hereinafter called ***quality alignment*** *in this paper). Meanwhile, the faces in the selected frames should also be* ***geometrically aligned*** *to the still faces offline well-aligned in the gallery. In this paper, we discover that the interactions among the three tasks–quality alignment, geometric alignment and face recognition–can benefit from each other, thus should be performed jointly. With this in mind, we propose a Coupling Alignments with Recognition (CAR) method to tightly couple these tasks via low-rank regularized sparse representation in a unified framework. Our method makes the three tasks promote mutually by a joint optimization in an Augmented Lagrange Multiplier routine. Extensive experiments on two challenging S2V datasets demonstrate that our method outperforms the state-of-the-art methods impressively.*

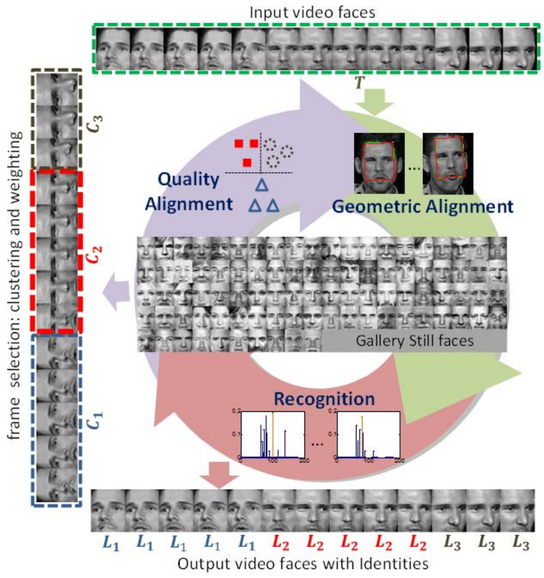
Input video faces
T
Quality
Alignment
Geometric Alignment
Gallery Still faces
Recognition
frame selection: clustering and weighting
C_3
C_2
C_1
L_1 L_1 L_1 L_1 L_1 L_2 L_2 L_2 L_2 L_2 L_3 L_3 L_3
Output video faces with Identities

Figure 1. Coupling Alignments with Recognition (CAR) framework. In our method, we jointly perform *geometric alignment*, *recognition* and *quality alignment* in a close loop to estimate the alignment parameters T, the identity labels L and the selecting confidences C for a video face sequence. Note that, all the face images in the top, left and bottom lines are from the same video.

1. Introduction

In recent years, there has been increasing interest in video-based face recognition for real world applications, especially person identification or retrieval in surveillance videos. Most of these works, e.g., [1, 2, 3, 4, 5], address the so called Video-to-Video (V2V) face recognition problem, in which query video sequences are matched against a set of target video sequences.

In real-world applications, however, a more practical scenario may be like this: the target set (or gallery) contains still images which are usually collected by a high quality digital camera under controlled environment, e.g., ID or driver license photo, thus of high resolution, in frontal view, with normal lighting and neutral expression. In contrast, the query (or probes) can be a video clip, which is taken under unconstrained environment by video surveillance cameras, thus usually of low image quality, such as low resolution, poor lighting, non-frontal poses, image blur and even serious misalignment. To differentiate it from the V2V face recognition problem, this scenario is specifically called the Still-to-Video (S2V) face recognition problem [6, 7].

To our best knowledge, only a few works [6, 7, 8] have studied the S2V face recognition problem. Most of them learned the relationship between the still images and video frames but did not directly handle bad quality frames, which very likely make the recognition perform badly. To deal with this problem, another kind of methods such as [9] is to first select the best quality frames and then integrate the recognition results of the selected frames. In this paper, we call the task of selecting good quality frames, with the most similar quality to that of still images, as *quality alignment*. As illustrated in the left part of Fig.1, the frames in the red box can be selected to match against the target faces, as these faces are in near frontal view. But, how to achieve accurate quality alignment forms the first challenge to attack for S2V face recognition scenario.

The second big challenge in S2V scenarios is the annoying geometric misalignment problem, which can lead to severe performance degradation [10]. Specifically, the problem arises because the faces taken from video can hardly be geometrically aligned accurately by existing methods (e.g., Active Shape Model (ASM) [11], Active Appearance Model (AAM) [12]), as the faces are generally of low resolution, probably with motion blurring and often taken under non-ideal lighting conditions. Furthermore, it is worth pointing out that here the "misalignment" means not only the mutual misalignment of the video frames but also their joint misalignment with the target faces. As an example, due to geometric misalignments, all the input video faces in Fig.1 are incorrectly identified by traditional recognition systems such as sparse representation-based classification (SRC) [13] etc. We call this alignment task *geometric alignment*, in contrast to the *quality alignment* mentioned above.

It is also worth noticing that the above two alignment problems are related. For example, intuitively, it is not necessary for us to geometrically align the target faces with those frame not selected by the quality aligning. On the contrary, the quality aligning results can be affected by the geometric (mis)alignment. Therefore, these two aligning process should be coupled in some way. Furthermore, it is clear that the above two types of alignments can heavily affect the recognition results. It has been well accepted that highly accurate recognition would be even impossible without precise alignment. However, the other side of the coin, i.e., correct recognition can feasibly lead to more accurate alignments, has been long neglected. In other words, alignment should not only simply precede recognition, but should also benefit from recognition. To put it in another way, the two kinds of alignments and recognition should be coupled together in a loop.

With above belief in mind, in this paper, we propose Coupling Alignments and Recognition (CAR) method which tightly couples the above two alignment tasks with *recognition* task, thus making them benefit each other in a unified loop, as shown in Fig.1. Specifically, we assume that if the faces in a video are accurately aligned with well-aligned gallery faces, they can be well represented as sparse linear combinations of the gallery faces with the same identity. This can connect and improve both of *geometric alignment* and *recognition* by simultaneously aligning and seeking sparse representations of video faces over gallery still faces. With better alignments and sparse representations, our proposed *quality alignment* can cluster and weight different quality frames more accurately. In addition, we also adopt low-rank prior that if video faces are in mild variations, a proper low-rank structure will exist. By incorporating the low-rank prior, each cluster of the same quality faces obtained by *quality alignment* can be jointly aligned and consistently represented as sparse linear combinations of gallery set, which can backward promote both *geometric alignment* and *recognition*. Consequently, in this close loop, our method iteratively aligns the video faces, identifies them and selects good frames, which can improve the three tasks mutually and finally corrects the initial possibly erroneous recognition decision.

The rest of the paper is organized as follows. Section 2 briefly reviews the related work. Section 3 details the proposed Coupling Alignments with Recognition (CAR) method for S2V face recognition. Section 4 presents our comprehensive experimental results on YouTube-S2V and COX-S2V datasets, followed by conclusions in Section 5.

2. Related Work

In this section,we briefly introduce the sparse representation for alignment and the low-rank representation for subspace segmentation.

2.1. Robust Alignment by Sparse Representation

Designed for *still images*, Robust Alignment by Sparse Representation (RASR) [10] simultaneously optimizes the alignment parameters and the sparse representation coefficients. Specifically, suppose that y is the probe face image which is misaligned and $A = [a_1, a_2, \dots, a_c]$ denotes the training dictionary with c subjects. To make a more accurate alignment, RASR assumes that the transformed image has a sparse representation [13] over A:

$$\min_{\alpha,\tau,e} \|\alpha\|_1 + \|e\|_1, \quad s.t. \quad y \circ \tau = A\alpha + e. \tag{1}$$

where α is the sparse coefficient vector, τ is the transformation and e is the residual error.

Since the model of RASR is a non-convex optimization problem, one of the difficulties is that it has many local minima that correspond to aligning y to different subjects which are not well-aligned. In this case, RASR turns to seek the best alignment of the test face from subject to subject:

$$\min_{\alpha_i,\tau_i,e_i} \|e_i\|_1, \quad s.t. \quad y \circ \tau_i = A_i\alpha_i + e_i. \tag{2}$$

where A_i is the matrix associated with subject i, τ_i is the transformation aligning y to subject i and e_i is the residual error of subject i in the transformation. Therefore, for a large-scale face database containing c subjects, Eq.(2) needs to be solved c times, making RASR have a high time consumption. To overcome this drawback, Misalignment-Robust Representation method (MRR) [14] firstly aligns the gallery images well offline and then directly solves its objective function in a global representation over the whole gallery images without concerning local minima.

2.2. Robust Subspace Segmentation by Low-Rank Representation

Low-Rank Representation (LRR) [15] effectively performs subspace segmentation with low-rank prior. Specifically, a set of data vectors $X = [X_1, X_2, \dots, X_c]$ are drawn from a union of c subspaces $\{S_i\}_{i=1}^{c}$, and X_i is the collection of n_i data vectors drawn from the i-th subspace S_i. Assuming each subspace owns a low-rank structure, they used the data X itself as the dictionary and find the lowest-rank representation that can represent the data vectors as linear combinations of the basis in the given dictionary:

$$\min_{Z} \|Z\|_*, \quad s.t. \quad X = XZ. \tag{3}$$

where $Z = [z_1, z_2, \dots, z_n]$ is the coefficient matrix with each z_i being the representation of x_i, $\|Z\|_*$ denotes the nuclear norm [16] of Z, i.e., the sum of the singular values of the matrix.

3. Proposed Method

In this section, we present Coupling Alignments with Recognition approach to jointly optimize three tasks—geometrically aligning faces (*geometric alignment*), performing *recognition* and selecting good quality frames (*quality alignment*)—in a unified framework. In the end, we also develop an efficient algorithm to solve it.

3.1. Formulation

The S2V face recognition problem matches low quality facial video frames against high quality gallery still faces. Let $Y = [y_1, y_2, \dots, y_n]$ be the misaligned probe video faces, $A = [A_1, A_2, \dots, A_c]$ be the gallery dictionary with c subjects, where A_i represents the well-aligned gallery still faces of the i-th subject. Formally, for the video faces Y, we want to estimate the alignment transformations T and the identity labels L simultaneously by the following:

$$\begin{aligned} \{\hat{T}, \hat{L}\} &= \arg\min_{T,L} \|Z\|_1 + \sum_{i=1}^{c} \|B_{S_i}\|_* + \|E\|_1, \\ s.t. \quad & Y \circ T = B + E, \quad B = AZ, \\ & S_i = \{j | L_j = i\}, \quad i = 1, 2, \dots, c, \quad j = 1, 2, \dots, n. \end{aligned} \tag{4}$$

where $T = [\tau_1, \tau_2, \dots, \tau_n]$ is the transformation matrix for the video faces, $Z = [z_1, z_2, \dots, z_n]$ are the sparse representation coefficient matrix of faces, $L_j = \arg\min_k \|y_j \circ \tau_j - A_k z_{jk}\|_2$ is the identity label of face y_j in recognition, S_i is the segment of faces with the i-th identity in recogntion, B's columns contain the sparse representations of video faces and $E = [e_1, e_2, \dots, e_n]$ is a matrix of the residual errors of video faces.

For joint geometric alignment and recognition, we adopt the sparse representation prior that if the alignments of video faces are accurate, they can be represented as good linear combinations of well-aligned gallery still faces. So, we need to seek an optimal set of deformations T for the video sequence Y simultaneously with their sparse representations over the gallery dictionary A. In this way, the sparse representations over gallery set make faces from video aligned with the gallery faces, thus geometrically align them more accurately. Meanwhile, the aligned video faces will obtain more accurate sparse representations in terms of the entire gallery set.

Furthermore, both better geometric alignment and recognition can facilitate quality alignment selecting good quality frames. Specifically, we assume that the video faces with the same recognition identity are similar in appearance. Under this assumption, video faces will be automatically clustered into different segments (i.e., $S_1, S_2, \dots, S_c$) according to the identities obtained by sparse representation-based classifier. Additionally, different clusters of faces will be weighted with different confidences, which are defined as:

$$C_{S_i} = \sum_{j=1, j \in S_i}^{|S_i|} exp\left(\frac{-\|e_j\|_1}{\sigma^2}\right) \tag{5}$$

where S_i is calculated in Eq.(4), $e_j = y_j \circ \tau_j - Az_j$, σ is empirically specified from the mean of $\|e\|_1$. Note that, this novel frame selection can also work for all Sparse Representation-based methods. With more accurate τ_j and z_j in the other two tasks, the reconstructed errors e_j will update the confidence C_{S_i} more accurately. Since the gallery dictionary only contains frontal faces, the reconstructed errors of faces in frontal cluster are often smaller than that of non-frontal ones. As a result, with more accurate clustering and weighting, quality alignment will select good quality frames with higher confidences.

Besides, the three tasks are also simultaneously coupled by low-rank prior, which assumes that since faces in one video vary continuously, they should own a good low-rank structure. When one video sequence has large inter-frame differences, better low-rank structures will be obtained in each of the individual clusters divided by quality alignment task. In these different video segments, the sparse representations of video faces are regularized by low-rank prior respectively to achieve more consistent linear combinations of gallery images and more accurate joint alignment of the faces with gallery images. Specifically, in our algorithm, we employ *coarse-to-fine* search strategy, which performs

```
Algorithm 1 Main Algorithm of the CAR method
INPUT: Gallery data matrix $A$, probe video sequence data matrix $Y$ and initial transformation $T$ of $Y$
1. WHILE not converged DO (outer loop)
2.     compute Jacobian matrices w.r.t transformations:
           $J_i = \frac{\partial}{\partial \zeta}\left(\frac{vec(I_i \circ \zeta)}{vec(\|I_i \circ \zeta\|_2)}\right)\Big|_{\zeta=\tau_i}, i = 1, \ldots, n;$
3.     warp and normalize the images:
           $Y \circ T = \left[\frac{vec(I_1 \circ \tau_1)}{vec(\|I_1 \circ \tau_1\|_2)}, \ldots, \frac{vec(I_n \circ \tau_n)}{vec(\|I_n \circ \tau_n\|_2)}\right];$
4.     set the segments at coarse search stage:
           $S_1 = \{1, \ldots, n\}, S_i = \phi, i = 2, \ldots, c$
5.     solve the linearized convex optimization: (inner loop)
           $\{\hat{T}, \hat{Z}\} = \arg\min_{T,Z} \|Z\|_1 + \sum_{i=1}^{c} \|B_{S_i}\|_* + \|E\|_1,$
           $s.t. \quad Y \circ T + J\Delta T = B + E, \quad B = AZ;$
6.     update transformations:
           $T = T + \Delta T^*;$
7.     update segments at fine search stage:
           $S_i = \{j | i = \arg\min_k \|y_j \circ \tau_j - A_k z_{jk}\|_2\}.$
8. END WHILE
9. Obtain the class label of probe video sequence with S2V classification method.
OUTPUT: Class label of the probe video sequence.
```

low-rank regularization on whole video at the *coarse* search stage and then conducts low-rank regularizations on different clusters divided by the identities at the *fine* search stage. More details will be described in subsection 3.2.

3.2. Optimization

The proposed model (4) involves nonlinearity problem of the constraint $Y \circ T = B + E$. Following recent work [10], in this paper, we adopt the iterative linearization scheme to solve the problem. After linearization, the optimization is relaxed as a linearized convex optimization problem, which can be efficiently resolved by Augmented Lagrange Multiplier Method [17].

3.2.1 Iterative linearization

Due to the complicated dependence of $Y \circ T$ on the transformations T, solving the nonlinearity of the constraint $Y \circ T = B + E$ is difficult. Nonetheless, we can approximate this constraint by linearizing about the current estimate of T when the change in T is small. After introducing the changes ΔT and the Jacobian $J = \frac{\partial}{\partial T} Y \circ T$ with respect to the transformation T, we write $Y \circ (T + \Delta T) \approx Y \circ T + J\Delta T$. Then, the nonlinearized optimization can be approximated as the following linearized formulation:

$$\begin{aligned}
\{\hat{T}, \hat{L}\} &= \arg\min_{T,L} \|Z\|_1 + \sum_{i=1}^{c} \|B_{S_i}\|_* + \|E\|_1, \\
s.t. \quad & Y \circ T + J\Delta T = B + E, \quad B = AZ, \\
& S_i = \{j | L_j = i\}, \quad i = 1, \ldots, c, \quad j = 1, \ldots, n.
\end{aligned} \tag{6}$$

The complete optimization procedure is summarized as Algorithm 1. In the input, the initial transformations of the probe video sequence could be the similarity transformations according to the automatically detected locations of eye centers. Steps from 1 to 8 are the outer loop, which iteratively linearizes the estimation of T and solves the convex function in Eq.(6). Specifically, step 2 and 3 compute the Jacobian matrices, warp and normalize the video faces. Step 4 sets only one segment (i.e., whole video sequence) when coarse searching, which is in the first couple of iterations of outer loop. Then step 5 conducts the inner loop, which solves the convex programming detailed in the following subsection. Next, we update the transformations at step 6. The segments are updated at Step 7 when fine searching, which is in remaining iterations of outer loop. Finally, we obtain class label of the probe video with S2V classification methods detailed in subsection 4.3.2.

3.2.2 Efficient Solution by Augmented Lagrange Multiplier Methods

In this work, we employ the Augmented Lagrange Multiplier (ALM) algorithm [17] to efficiently solve the convex function at step 4 in Algorithm 1. The basic idea of the ALM method is to search for a saddle point of the augmented Lagrangian function instead of directly solving the original constrained optimization problem. For our problem (6), the augmented Lagrangian function is given by:

$$\begin{aligned}
\ell_u(B, E, \Delta T, X) = \sum_{i}^{m} \|B_{S_i}\|_* + \langle X, h(B, E, \Delta T)\rangle \\
+ \lambda\|E\|_1 + \frac{\mu}{2}\|h(B, E, \Delta T)\|_F^2
\end{aligned} \tag{7}$$

where $h(B, E, \Delta T) = Y \circ T + J\Delta T - B - E$, X is a Lagrange multiplier matrix, μ is a positive scalar, $\langle \cdot \rangle$ denotes the matrix inner product, and $\| \cdot \|_F$ denotes the Frobenius norm.

For appropriate choice of Lagrange multiplier matrix X and sufficiently large constant μ, it can be shown that the augmented Lagrangian function has the same minimizer as the original constrained optimization problem [18]. To estimate both the Lagrange multiplier and the optimal solution, it is typical to iteratively minimize the Lagrangian function approximately by minimizing the function against the three

Algorithm 2 Algorithm of the CAR's **inner loop**

```
INPUT: $(B^0, S, E^0, \Delta T^0, A)$
1. WHILE not converged DO
2.    $B^{k+1} = Y \circ T + J\Delta T + \frac{1}{\mu^k} X^k - E^k$;
3.    $B^{k+1} = A(A^T A + \lambda I)^{-1} A^T B^{k+1}$;
4.    $(U_i, \Sigma_i, V_i) = svd(B_{S_i}^{k+1}), i = 1, 2, \ldots, c$;
5.    $B_{S_i}^{k+1} = U_i \Gamma_{\frac{1}{\mu^k}}[\Sigma_i] V_i^T, i = 1, 2, \ldots, c$;
6.    $E^{k+1} = \Gamma_{\frac{\lambda}{\mu^k}}[Y \circ T + J\Delta T + \frac{1}{\mu_k} X^k - B^{k+1}]$;
7.    $\Delta T^{k+1} = J^{\dagger}(B^{k+1} + E^{k+1} - Y \circ T - \frac{1}{\mu^k} X^k)$;
8.    $X^{k+1} = X^k + \mu^k h(B^{k+1}, E^{k+1}, \Delta T^{k+1})$.
9. END WHILE
OUTPUT: Solution $(Z^*, B^*, E^*, \Delta T^*)$ to problem (6).
```

unknowns B, E, T one at a time:

$$\begin{aligned} B^{k+1} &= \arg\min_{B} \ell_{u^k}(B, E, \Delta T, X^k) \\ E^{k+1} &= \arg\min_{E} \ell_{u^k}(B, E, \Delta T, X^k) \\ \Delta T^{k+1} &= \arg\min_{\Delta T} \ell_{u^k}(B, E, \Delta T, X^k) \\ X^{k+1} &= X^k + \mu^k h(B^{k+1}, E^{k+1}, \Delta T^{k+1}) \end{aligned} \tag{8}$$

Since each step of the above iteration involves solving a convex program, the problem can be solved efficiently by a single step. To spell out of the solutions, let $\Gamma_\alpha[x] = sign(x) \cdot \max\{|x| - \alpha, 0\}$ be the *soft-thresholding* or *shrinkage operator* for scalars, where $\alpha \geq 0$. Using the shrinkage operator, we can rewrite the solution to each of (8) at steps 2-8 of Algorithm 2. Following [19], step 3 calculates the collaborative representations of video faces over dictionary. The $svd(\cdot)$ at step 4 denotes the Singular Value Decomposition operator, and $J\dagger$ at step 7 denotes the Moore-Penrose pseudoinverse of J. In our experiment, the algorithm always converges to the optimal solution to the problem (6), and does so significantly faster than other alternative convex optimization methods.

4. Experiments

In this section, we present extensive experiments to demonstrate the effectiveness of our proposed Coupling Alignments with Recognition (CAR) algorithm in terms of both alignment accuracy and recognition performance. In this work, we conduct evaluations on two S2V datasets: YouTube-S2V collected from the YouTubeDB [20] and COX-S2V [7], which are detailed in following subsection.

For alignment, we compare our approach with one of the state-of-the-art blind joint alignment algorithm RASL [21]. Besides, the alignment results of recently proposed simultaneous alignment and recognition method MRR [10] is also shown. For S2V recognition, we compare our CAR algorithm with the following methods: (1)SRC/CRC: directly input the original data into SRC algorithm [13] or CRC algorithm [19]; (3)A-SRC/A-CRC: feed the data with alignment of RASL into SRC or CRC; (5)MRR: jointly aligns and identifies the video faces frame by frame.

4.1. Evaluation Datasets

The YouTube-S2V dataset is collected by us from the YouTube Faces DB [20], in which the videos are downloaded from YouTube. To design a S2V scenario, we select the videos of 100 different subjects from it while the still images of these subjects are collected by Google Image Search. As it is not easy to collect more than one high quality frontal face images for ALL the subjects, to keep uniform for all the subjects, we included only 1 frontal face image for each person in the gallery, as in COX-S2V [7]. For evaluation of S2V face recognition, we design an unsupervised scenario, which uses the still images for gallery and the videos for probe without any training set.

The COX-S2V [7] dataset contains 1,000 subjects, with each subject 1 high quality photo and 4 video clips captured by normal camcorders. The four kinds of test video sequences are in different qualities: Both video1 and video3 sequence contain low resolution faces (around 16 × 20) while the faces in video2 and video4 are of relatively higher resolution (around 48 × 60); The face poses in video1 and video2 are of nearly-controlled while the poses in video3 and video4 are unconstrained. In the protocol, the images and videos of 300 persons are used for training and the rest 700 person's corresponding data are used for testing. In the test, the still images are enrolled in the gallery while the video sequences are contained in the probe. To design an unsupervised case on COX-S2V, we also conduct experiments on it without the training set.

4.2. Experimental Settings

In our experiments, a commercial face detection SDK OKAO[1] was employed to detect faces and locate eyes in both still images and videos. For the YouTube-S2V dataset, the size of the normalized face image is set to 48×60, and the coordinates of the eye centers are $(12, 27)$ and $(36, 27)$ respectively. For the COX-S2V dataset, the size of the normalized face image is set to 96×120, and the coordinates of eye centers are set to $(24, 54)$ and $(72, 54)$ respectively.

The methods RASL (in A-SRC, A-CRC) and MRR are performed using codes from the original authors. As the original work of RASL, the method's default parameters are used in our experiments. The initial transformation of each face image is calculated according to the automatically detected eye center positions and the type of transformation in RASL is the similarity transformation, which is also used in both MRR method and our method CAR. For MRR algorithm, we use the first η columns of U and set

[1] http://www.omron.com/r_d/technavi/vision/okao/detection.html

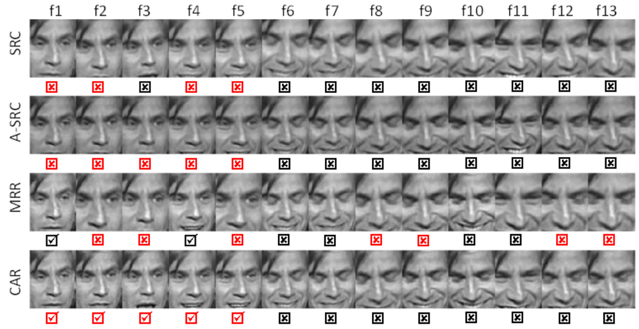

Figure 2. Rows from top to bottom show the alignment and identification results of SRC(original faces), A-SRC(RASL), MRR and CAR respectively on one video sequence from YouTube-S2V. The tick indicates the face is correctly identified, and the red box indicates the face frame is selected in quality alignment.

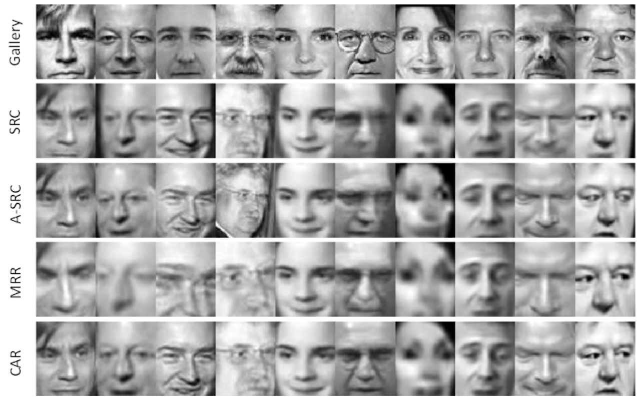

Figure 3. Gallery faces and average faces from videos before and after alignment. SRC shows the average of original video faces from a face detector using its quality alignment result; and A-SRC(RASL), MRR, CAR show the average faces after their respective alignments.

$\eta = 25$. The regularized parameter λ in sparse representation and collaborative representation is set to 3 and 0.05 respectively. The other parameters of MRR are setted as follows: $\eta_1 = 25, \eta_2 = 40, S = 25$. The iteration number of seeking the optimal solution in RASL, MRR is set to 25. The iteration number of coarse searching in CAR is 13, while that of fine searching is 12.

4.3. Evaluation results on S2V alignment and recognition

In our CAR algorithm, the tasks of alignments and recognition are tightly coupled. However, to facilitate the comparison with conventional alignment and recognition approaches respectively, we will present the results for alignments and recognition respectively in the subsection.

4.3.1 Evaluation of Alignment Accuracy

We illustrate results on the YouTube-S2V database to evaluate the alignment accuracies of RASL, MRR and our method CAR. Before alignment, we obtain an initial estimate of the transformation in each image using the off-the-shelf face detector. As a blind alignment method, original RASL jointly aligns the video faces without considering the gallery still faces. To align with the gallery images, in this experiment, RASL is used to add all the gallery still images into the video sequence and jointly align the still faces and video faces together. As a simultaneous alignment and recognition method designed for still images, we use MRR to align the video faces frame by frame in each video clip. In contrast, our algorithm CAR jointly align the video faces mutually with the gallery still images.

Fig.2 shows alignment and recognition results of one video sequence with 13 selected frames from the YouTube-S2V dataset. As shown in Fig.2, the misalignments of input video faces are very serious: the eyes of most faces are not in horizontality, the face scales are not the same, and most ones do not have the whole mouth. Although RASL jointly aligns the eyes of video faces in the same horizontality, most aligned faces have partial mouth. MRR aligns several video faces accurately, but it still makes some faces in a wrong scale, such as f2,f3,f5. In contrast, except faces f11-f13 owning exaggerated facial expressions, our method CAR aligns most of faces including f2,f3,f5 to the gallery still image. In addition, faces in red box are the selected ones in quality alignment. The results manifest our method CAR can also achieves more accurate quality alignment.

Since there is no ground truth for this dataset, we verify performances of involved method visually by plotting the average faces before and after geometric alignment and quality alignment. Fig.3 shows the gallery faces and the mean faces of videos from 10 subjects in YouTube-S2V. For example, in Fig.3, the first average face of SRC is the mean of the faces with red box (i.e., f1,f2,f4,f5) in SRC row of Fig.2. Note that the average faces after CAR's alignment are more clear and aligned with gallery images more accurately than those of other methods. This result suggests that our method CAR achieves the improved geometric alignment in unconstrained S2V scenario.

This can be explained by that, via jointly exploiting the sparse representation prior and low-rank prior, our method demonstrates much more robustness in aligning the video faces. On one hand, the sparse representation over gallery dictionary makes the video faces aligned to the still images, which are well-aligned. But, in the S2V case, several faces are not robustly aligned only with the sparse representation. Consequently, on the other hand, the low-rank prior facilitates those more easily aligned video faces to correct the alignments of the others in the same video sequence. For example, as shown in Fig.2, the bad aligned faces f2,f3,f5

Table 1. Unsupervised S2V face recognition results (rank-1 recognition rate (%)) on YouTube-S2V(Y) and COX-S2V(Ci) datasets.

Methods	Min					Voting					C-Voting				
	Y	C1	C2	C3	C4	Y	C1	C2	C3	C4	Y	C1	C2	C3	C4
SRC	8.62	15.43	40.57	1.86	14.57	8.62	14.29	38.43	2.00	14.57	10.78	15.57	42.29	2.86	18.71
CRC	7.76	14.86	41.00	3.00	14.43	8.62	14.57	38.86	3.86	17.71	10.34	14.43	43.57	4.14	19.71
A-SRC	19.82	20.57	37.71	4.14	16.43	23.27	21.71	37.00	4.00	17.14	26.29	22.14	39.00	4.57	18.29
A-CRC	20.26	16.43	39.86	2.57	16.43	25.00	18.71	41.14	4.14	18.71	29.74	19.43	41.29	4.00	19.43
MRR	21.55	25.71	42.71	3.71	12.71	25.43	28.00	43.71	4.57	13.29	28.45	26.43	44.14	3.57	13.57
CAR	**24.57**	**38.57**	**52.57**	**5.43**	**21.00**	**30.17**	**42.14**	**54.14**	**9.14**	**25.14**	**36.21**	**43.42**	**55.00**	**10.71**	**28.86**

Table 2. Supervised S2V face recognition results (rank-1 recognition rate (%)) on COX-S2V (Ci) dataset.

Methods	C1	C2	C3	C4
LDA	47.57	68.28	20.00	49.58
PaLo	52.43	73.00	22.00	56.71
SRC	55.43	71.14	19.86	52.43
CRC	58.57	68.51	22.14	51.29
A-SRC	40.86	63.57	11.00	40.43
A-CRC	39.43	64.71	12.29	40.86
MRR	64.43	68.71	26.14	53.29
CAR	**72.57**	**77.29**	**35.43**	**61.43**

in MRR are correctly aligned by f1,f4 in CAR. Due to better geometric alignment of CAR, the faces f2,f3,f5 are also identified correctly.

4.3.2 Evaluation of S2V Face Recognition

The S2V recognition involves defining the similarity between video sequence and gallery images and determining which strategy is used for classification. In this experiment, we adopt the image-set based similarity $D_{jk} = \|y_j \circ \tau_j - A_k z_{jk}\|_2$, which is used in SRC [13], and the following three different strategies for S2V classification:

- Min: $id_Y = \arg\min_k D_{jk}$, where $j = 1, \ldots, n$;
- Voting: $id_Y = \arg\max_i \|S_i\|$, where $S_i = \{j | i = \arg\min_k D_{jk}\}$;
- C-Voting: voting strategy with confidences obtained by quality alignment: $id_Y = \arg\max_i C_{S_i}$, where C_{S_i} is defined in Eq.(5), we empirically set $\sigma = 0.4$ according to the mean of error $\|e\|_1$.

For S2V face recognition on YouTube-S2V and COX-S2V, we first evaluate the comparative methods in unsupervised case, where additional training set are not used for obtaining the discriminant information. Table 1 summarizes the S2V recognition results on the two S2V datasets. The columns of Y and Ci represent the testing on YouTube-S2V and video i (as detailed in section 4.1) of COX-S2V respectively. Before comparing involved methods, we need to compare the three different classification strategies. The experimental results demonstrate that the C-Voting is better than other two strategies for S2V recognition. This is because that the C-Voting with quality alignment is more suitable for S2V scenario. In the CAR's result of Fig.2, the f1-f5 are identified as a right ID while f6-f11 are recognized as a wrong ID. In this case, although the Voting will give the wrong final identification, the C-Voting will identify the video correctly by favoring f1-f5 with higher confidences.

Then, we conclude the results of different methods: Due to the blind alignment preprocess of RASL, A-SRC and A-CRC performs slightly better than the original methods SRC and CRC. By simultaneously aligning and recognizing faces frame by frame, MRR works better than RASL. In contrast, our CAR algorithm outperforms the other methods remarkably. This is because the proposed method CAR improves the recognition by both more accurate geometric alignment and better frame selections in quality alignment: the geometric alignment generates better sparse representations over gallery set by jointly aligning the video faces more accurately. In different clusters divided by quality alignment, low-rank regularizations are conducted on the sparse representations of video faces to make the sparse representation-based classification more robust.

In the supervised experiment, we use the additional training set of COX-S2V to train LDA model to extract discriminant features for both still images and video frames. In training, for A-SRC, A-CRC, MRR and CAR, the training images are all aligned in advance by the corresponding alignment methods. In test stage, we still use gray feature before the identification and LDA feature at the final recognition. Due to space limitation and in order to fairly compare with the benchmark work [7] on COX-S2V, we use Min strategy in the supervised scenario. As shown in Table 2, comparing with the original method LDA and the method PaLo [7] on COX-S2V, our method CAR significantly outperform them with average gains of 15.32% and 10.65% respectively. The superiority of CAR demonstrates that the face misalignment in videos is indeed one of the most importance challenges in S2V scenario. Besides, CAR again performs much better than other methods in the supervised S2V face recognition. This supervised case also suggests that the idea coupling alignments and recognition is utterly desirable for S2V face recognition.

4.4. Complexity analysis

The main time complexity of of our method is solving the linearized convex optimization (i.e., Equ. (7)) by Augmented Lagrange Multiplier (ALM). Our ALM solver mainly contains SVD operation, whose complexity is $O(m^2n + n^3)$, where m is the dimension of feature and n is the number of video frames. As $m \gg n$ when n is reduced to the number of frames in one cluster at the fine-searching steps, the complexity is $O(m^2)$.

Here we take the experiment on YouTube-S2V dataset as example: on a 2.93 GHz Intel(R) Core(TM) i7 CPU machine, the MATLAB implementation of our approach requires about 160 seconds to simultaneously align and identify a video sequence containing about 157 frames. This speed is similar to that of MRR (about 144 seconds), which to our best knowledge is the most related and by far the fastest method (about 72 times speedup to RASR [20]) solving the problem addressed in this paper.

5. Conclusion

In this paper, we first studied the mutual influence among geometric alignment, quality alignment and recognition in the S2V face recognition scenario. The mutual promotions among the three tasks inspired us to propose a method Coupling Alignments and Recognition (CAR), which tightly combines the three tasks by making full use of sparse representation prior and low-rank prior. As far as we know, the proposed method is the first attempt to jointly perform and simultaneously improve the above three tasks in a unified framework for the S2V application. By jointly considering the three interactive tasks, our algorithm has demonstrated significant superiority over those methods treating them separately through extensive experiments on two challenging datasets YouTube-S2V and COX-S2V.

Acknowledgements. The work is partially supported by Natural Science Foundation of China under contracts nos. 61025010, 61222211 and 61173065; and Beijing Natural Science Foundation (New Technologies and Methods in Intelligent Video Surveillance for Public Security) under contract no. 4111003.

References

[1] X. Liu and T. Cheng. Video-based face recognition using adaptive hidden markov models. In *CVPR*, 2003.

[2] T.K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE T-PAMI*, 29(6):1005–1018, 2007.

[3] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, 2010.

[4] Y. Hu, A.S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *CVPR*, 2011.

[5] R. Wang, S. Shan, X. Chen, Q. Dai, and W. Gao. Manifold-manifold distance and its application to face recognition with image sets. *IEEE T-IP*, 21(10):4466–4479, 2012.

[6] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *CVIU*, 91(1):214–245, 2003.

[7] Z. Huang, S. Shan, H. Zhang, H. Lao, Alifu. Kuerban, and X. Chen. Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on COX-S2V dataset. In *ACCV*, 2012.

[8] S. Biswas, G. Aggarwal, and P.J. Flynn. Pose-robust recognition of low-resolution face images. In *CVPR*, 2011.

[9] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B.C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *CVPR Workshops*, 2011.

[10] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Towards a practical face recognition system: robust alignment and illumination by sparse representation. *IEEE T-PAMI*, 34:372–386, 2012.

[11] T. Cootes, G. Edwards, and C. Taylor. Active shape models—'smart snakes'. In *BMVC*, 1992.

[12] T. Cootes and C. Taylor. Active appearance models. *IEEE T-PAMI*, 23:681–685, 2001.

[13] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE T-PAMI*, 31:210–227, 2009.

[14] M. Yang, L. Zhang, and D. Zhang. Efficient misalignment-robust representaion for real-time face recognition. In *ECCV*, 2012.

[15] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, 2010.

[16] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002.

[17] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Fast l_1-minimization algorithms and application in robust face recognition. *Technical Report UCB/EECS-2010-13, UC Berkeley*, 2010.

[18] D.P Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

[19] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation which helps face recognition? In *ICCV*, 2011.

[20] L. Wolf, H. Tal, and M. Itay. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011.

[21] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. RASL: Robust alignement by sparse and low-rank decomposition for linearly correlated images. In *CVPR*, 2010.