Batch Algorithm with Additional Shape Constraints for Non-Rigid Factorization

Yuan Ren Loke and Ranganath Surendra
Department of Electrical and Computer Engineering
National University of Singapore
4 Engineering Drive 3, Singapore 117576
{g0500069,elesr}@nus.edu.sg

Abstract

Recently, recovery of non-rigid structure by the factorization algorithms have received attention in the literature. The factorization algorithm decomposes the feature points over the given image sequence into motion of the camera and 3D shape bases. The non-rigid structure can be represented by the linear combination of the 3D shape bases. Although the closed-form solution of the non-rigid factorization algorithm is proven, the algorithm is sensitive to noise. In this paper, we propose a batch algorithm to recover multiple non-rigid structures from subsets of the data. Then, we introduce a set of non-linear shape constraints to optimize the recovered non-rigid structures. Synthetic data and real data were used in the experiments. The experimental results showed that the new factorization algorithm gives significant improvement than the original algorithm. With noisy data, the new algorithm is more robust and more accurate in recovering non-rigid structure.

1 Introduction

Recovering 3D structure from a sequence of images is one of typical interest topics in the computer vision community. In the past two decades, factorization algorithms have been widely applied to structure from motion (SFM) problems. It was first introduced to reconstruct rigid structure under arbitrary motion by Tomasi and Kanade [11]. Basically, the factorization algorithm for SFM decomposes the image feature tracks (*measurement matrix*) into motion of the camera and the 3D shape matrix via Singular Value Decomposition (SVD) and rank theorem. However, it is an ill-conditioned problem. Their linear transformations also yield valid motions and bases. Therefore, it is not possible to recover structure from the image sequence without some prior knowledge. Additional constraints such as orthogonality of rotation matrix are required to recover the structure.

Generally, orthographic camera model is chosen as the camera model for the factorization algorithm because it is a good approximation to the perspective camera model when the reconstructed target is far from the camera and the depth variation within the target is relatively small. [10] and [8] also proposed extended factorization algorithms for perspective and paraperspective models, respectively.

Recently, recovery of different kinds of structures such as multiple linearly moving objects [7], articulated objects [14], model based non-rigid objects [3], [1], [12], [13] are

BMVC 2007 doi:10.5244/C.21.95

reported. Model based non-rigid object recovery is attractive because many interesting non-rigid objects in nature such as human face can be represented by models. Reconstructing 3D human faces is very useful in face recognition. Compared to 2D face images, 3D face are invariant to pose changes. The pose changes significantly affect the performance of face recognition algorithms. Therefore, we can use non-rigid factorization to decompose the pose and deformation of the non-rigid structure from a image sequence.

To model the deformation of these non-rigid objects, the weighted combination of basis shapes has been applied in non-rigid SFM [3]. Using this model, Jing Xiao et al. [13] showed a closed-form solution for non-rigid SFM with rotation constraints and basis constraints. The solution is exact only when the data is noise free. The method does not work satisfactorily with noisy data [2].

In this paper, a batch algorithm and a non-linear shape constraint optimization are proposed to improve the existing closed-form solution under noisy environments. The batch algorithm partitions the matrix and recovers 3D structures from each partition separately. Then we apply the optimization algorithm to refine the closed-form solution of each partition based on shape constraints. Qualitative and quantitative evaluation showed that the new algorithm gives more robust and more accurate results compared to the original factorization method for both rigid and non-rigid structure.

2 Overview of Factorization Algorithm for Non-rigid SFM

Here the camera model is assumed to be the weak perspective projection model. We also assumed that the motion is non-degenerate. Let the 2D image coordinates of P feature points over F frames denoted as $\mathbf{W} = \{\mathbf{w}_{fp} = (u_{fp}, v_{fp}) | f = 1, \ldots, F, p = 1, \ldots, P\}$, the $2F \times P$ *measurement matrix*:

$$\mathbf{W} = \begin{bmatrix} u_{11} & \ldots & u_{1P} \\ v_{11} & \ldots & v_{1P} \\ \vdots & u_{fp} & \vdots \\ \vdots & v_{fp} & \vdots \\ u_{F1} & \ldots & u_{FP} \\ v_{F1} & \ldots & v_{FP} \end{bmatrix} \tag{1}$$

The camera projection matrix is written as:

$$\mathbf{R}_f = \begin{bmatrix} r_{f1} & r_{f2} & r_{f3} \\ r_{f4} & r_{f5} & r_{f6} \end{bmatrix} \qquad f \in \{1, \ldots, F\} \tag{2}$$

The non-rigid structure is represented by a linear combination of K 3D shape bases. Let $\mathbf{S}_f = \{\mathbf{s}_{fp} = (x_p, y_p, z_p) | p = 1, \ldots, P\}$ denote the 3D non-rigid structure of the f^{th} frame. Let $\mathbf{B} = \{\mathbf{b}_k = (x_{kp}, y_{kp}, z_{kp}) | k = 1, \ldots, K, p = 1, \ldots, P\}$ denote as the 3D shape bases. Then, the 3D non-rigid structure in each frame can be represented as:

$$\mathbf{S}_f = \sum_{k=1}^{K} c_{fk} \mathbf{b}_k \qquad f \in \{1, \ldots, F\} \tag{3}$$

where c_{fk} are the weights. Then, $\mathbf{W} = \mathbf{MB} + \mathbf{T}$ where $\mathbf{M}$ is a $2F \times 3K$ motion matrix, $\mathbf{B}$ is a $3K \times P$ 3D structure matrix and $\mathbf{T}$ is a $2F \times 1$ translation vector. When K=1, the structure is rigid. The motion matrix is the product of the weighting coefficients and the corresponding camera projection matrices. We can write this as

$$\mathbf{M} = \begin{bmatrix} c_{11}\mathbf{R}_1 & \dots & c_{1K}\mathbf{R}_1 \\ \vdots & c_{fk}\mathbf{R}_f & \vdots \\ c_{F1}\mathbf{R}_F & \dots & c_{FK}\mathbf{R}_F \end{bmatrix} \tag{4}$$

The translation vector can be obtained by computing the mean of the P feature points. The *registered measurement matrix*, $\hat{\mathbf{W}}$ is given by subtracting $\mathbf{T}$ from $\mathbf{W}$. The world origin now is placed at the centroid of the feature points, i.e.

$$\frac{1}{P}\sum_{p=1}^{P} \mathbf{w}_{fp} \qquad \forall f \in \{1,\dots,F\} \tag{5}$$

When the data is noiseless, the rank of $\hat{\mathbf{W}}$ is $3K$. Applying SVD, $\hat{\mathbf{W}}$ can be decomposed into a motion matrix, $\hat{\mathbf{M}}$ and a 3D basis matrix, $\hat{\mathbf{B}}$. However, it is only up to an arbitrary $3K \times 3K$ invertible transformation, $\mathbf{G}$. The exact motion matrix, $\mathbf{M}$ and 3D basis matrix, $\mathbf{B}$ can be written as:

$$\begin{aligned} \mathbf{M} &= \hat{\mathbf{M}} \cdot \mathbf{G} \\ \mathbf{B} &= \mathbf{G}^{-1} \cdot \hat{\mathbf{B}} \end{aligned} \tag{6}$$

The *corrective transformation matrix*, $\mathbf{G}$ is compound of K $3K \times 3$ matrix, $\mathbf{G}_k$. Then, $\mathbf{Q}_k = \mathbf{G}_k\mathbf{G}_k^T$. Computing the $\mathbf{Q}_k$ requires additional constraints. We have

$$\hat{\mathbf{M}}\mathbf{Q}_k\hat{\mathbf{M}}^T = \begin{bmatrix} c_{1k}\mathbf{R}_1 \\ \vdots \\ c_{1k}\mathbf{R}_F \end{bmatrix} \begin{bmatrix} c_{1k}\mathbf{R}_1 & \dots & c_{1k}\mathbf{R}_F \end{bmatrix} \tag{7}$$

Since rotation matrices are orthonormal, we have $\mathbf{R}_i\mathbf{R}_i^T = \mathbf{I}_{2\times 2}$. In [13], it was showed that using only these rotation constraints is insufficient to uniquely determine $\mathbf{Q}_k$. Thus, they also assume the first K images to be basis images. The corresponding weighting coefficients are then

$$c_{ij} = \begin{cases} 1 & \text{when } i = j \\ 0 & \text{when } i \neq j \end{cases} \tag{8}$$

We can now obtain a closed-form solution for each $\mathbf{Q}_k$ by optimizing the rotation and basis constraints. For the details of proof, the reader is referred to [13].

3 Batch Algorithm Using Matrix Partitioning

In practice, a large number of frames from video sequence are available, and using all the frames in SVD algorithm to minimize $\|\mathbf{W} - \mathbf{MB}\|_F$ may bring no advantage, firstly, because there is a large amount of redundancy in the video frames (this is just increasing the computational cost). and secondly, minimizing $\|\mathbf{W} - \mathbf{MB}\|_F$ does not guarantee that

the recovered structure is optimal. The solutions of the motion matrix $\mathbf{M}$ and the bases $\mathbf{B}$ also depend on the constraints we used on solving the corrective transformation matrix, $\mathbf{G}$.

Hence, we introduce a batch algorithm where a registered measurement matrix is partitioned into N submatrices. Then, the closed-form solution method is applied to each separately. This yields N estimates instead of a single estimate for the structures from a large number of frames. We hence expect that the proposed algorithm will improve the confidence in the result. We then propose to use these in a shape constrained non-linear optimization technique to find the best shape estimate.

Let $\Omega_i \subset \{1,\ldots,F\}$, $i = 1,\ldots,N$ be a subset of frame indexes. Then, let $\mathbf{W}_{\Omega_i} = \{(u_{fp}, v_{fp}) | f \in \Omega_i, p = 1,\ldots,P\}$ denote a row subspace of the matrix, where $|\Omega_i| \geq \max(\frac{K^2+K}{2}, 3K)$. The union of all subsets Ω_i contains all the elements of $\{1,\ldots,F\}$. All subsets are disjoint. Hence, the information in every frame is used for recovery of the structure.

Here, we assume that K is known. The set of K basis images which give the smallest condition number is the set of the most independent basis images. Thus, we can selected them as the K basis images.

Since the rank of $\hat{\mathbf{W}}_{\Omega_i}$ has to be at least $3K$, the number of frames in each partition can be determined in such a way that reasonable amount of the energy of $\hat{\mathbf{W}}_{\Omega_i}$ remains in the first $3K$ eigen-subspaces. Then each $\hat{\mathbf{W}}_{\Omega_i}$ can be decomposed by the non-rigid factorization algorithm discussed in Section 2 as:

$$\mathbf{W}_{\Omega_i} = \mathbf{M}_{\Omega_i}\mathbf{B}_i \qquad i = 1,\ldots,N \tag{9}$$

The recovered structures are exact for noiseless data.

When $K = 1$ (rigid case), the motion matrix $\mathbf{M}$ and $\mathbf{B}$ are simplified as rotation matrix $\mathbf{R}$ and the rigid structure matrix $\mathbf{S}$. When $K \geq 2$ (non-rigid case), we do not only need to recover the bases $\mathbf{B}$, but also the weighting coefficients in the motion matrix $\mathbf{M}$ for recovering the 3D structure. $\mathbf{M}$ can be obtained as

$$\mathbf{M}_i = \mathbf{W}\mathbf{B}_i^{+} \qquad i = 1,\ldots,N \tag{10}$$

where $\mathbf{B}_i^{+}$ is the pseudo-inverse of $\mathbf{B}_i$. Since the rotation matrix $\mathbf{R}_f$ is orthonormal, $||\mathbf{R}_f|| = 1$. The corresponding coefficients for each frame can be easily extracted out from motion matrix.

Let N sets of the estimated structures of the f^{th} frame denote as $\{\tilde{\mathbf{S}}_f\}_i$. Given the 3D shape bases $\mathbf{B}_i$ and the corresponding coefficients, each recovered structure can be computed by (3). Since each set of the recovered structures, $\{\tilde{\mathbf{S}}_f\}_i$ is independently estimated from the corresponding $\mathbf{W}_{\Omega_i}$, the reference coordinate systems of each two sets of the recovered structures are different up to a 3×3 orthonormal transformation. The orthonormal transformation can be obtained by applying Procrustes method.

4 Non-linear Shape Constraint Optimization

Here, we introduce an objective function which is optimized to enforce non-linear shape constraints and estimate the best recovered structure $\mathbf{S}_f$ from the set of estimated struc-

tures $\{\tilde{\mathbf{S}}_f\}_i$ from each partition. It is given as:

$$\min \sum_{n=1}^{N} \sum_{i=1}^{P} \sum_{j=1}^{P} \|s_{fi} s_{fj}^T - \tilde{s}_{fin} \tilde{s}_{fjn}^T\|^2 \qquad \forall f \in \{1, \ldots, F\} \tag{11}$$

where N is the number of partitions. This optimization minimizes the inner products of every two feature points. In other words, we are optimizing the errors in the lengths and the mutual angles of the feature points, so we named it *metric optimization*. The metric optimization plays a role in structure refinement of the factorization method. A general-purpose quasi-Newton method [4],[5],[6],[9] is used to find the optimum solution of (11).

The initialization is critical for non-linear optimization problems. To avoid the solution of the metric optimization from being trapped at an unsuitable local minimum, we choose the least mean square of $\{\tilde{\mathbf{S}}_f\}_i$ as the initial value for the metric optimization. In the experiments discussed in the following section, we show that the metric optimization gives more robust and better results than the original algorithm.

Our proposed algorithm is summarized as follows:

1. Partition the measurement matrix $\mathbf{W}$ into N submatrices.
2. Choose the K basis images from each subset based on their condition numbers. The set of the K basis images with the smallest condition number is the set of the most independent basis images.
3. Apply non-rigid factorization algorithm proposed by Jing Xiao et al. [13]
4. Extract the weight c_{fk} from the motion matrix $\mathbf{M}$.
5. Compute the structures by Eq. (3).
6. Optimize the estimated structures obtained in Step 5 by the objective function in Eq. (11).

5 Experiments

We evaluated the proposed factorization algorithm with metric optimization quantitatively and qualitatively on synthetic data and facial expression images, respectively. In the quantitative evaluation, our approach was applied on rigid and non-rigid synthetic data sets. In the qualitative evaluation, a set of human face expressions was used to examine the performance of our approach. The results are presented below.

5.1 Quantitative Evaluation on Synthetic Data

In this section, three approaches were evaluated on synthetic data. The first approach is Jing Xiao et al's [13] non-rigid factorization algorithm. The second approach applies the batch algorithm to estimate the 3D structures from each partition. The optimum structure is the mean of the estimated 3D structures which was the smallest mean square distance to the 3D estimated structures. The third approach is the batch algorithm with metric optimization. Two experiments were carried out to examine the performance of the algorithms.

In the first experiment, 15 rigid object datasets with Gaussian white noise were generated. The strength level of noise is defined as $\frac{\|\mathbf{noise}\|}{\|\mathbf{W}\|}$. Each dataset has 50 3D feature points and 100 frames with random projection matrices. A 200×50 measurement matrix $\mathbf{W}$ represented the image feature tracks. In the second experiment, 5 non-rigid object datasets formed by 3 shapes bases were generated. Each dataset has 25 3D feature points and 203 random projection matrices. A 406×25 measurement matrix $\mathbf{W}$ represented the image feature tracks. For the non-rigid dataset, Gaussian white noise was added at strength levels of 5%, 10% and 20% to evaluate the performance of the algorithms.

To make the experiments comparable, all the synthetic datasets were partitioned into 10 subsets and batch algorithm of section 3 was applied. For rigid case, each subset contains 50 3D feature points and 10 random projection matrices. For non-rigid case, each subset contains 25 3D feature points and 13 random projection matrices (3 basis images + 10 non-basis images). They formed 10 smaller measurement matrices $\mathbf{W}_i$. Then we applied non-rigid factorization algorithm on each $\mathbf{W}_i$ to recover 3D structures. Metric optimization is applied on these 3D estimated structures by quasi-Newton optimization algorithm.

For rigid case, the relative error measurement, $\frac{1}{P}\sum_{p=1}^{P}\frac{\|\mathbf{b}_p-\mathbf{b}_p^{truth}\|}{\|\mathbf{b}_p^{truth}\|}$ was evaluated for examining the performance of our approach. For non-rigid case, the mean of the relative errors between the optimal structure and the ground truth, $\frac{1}{PF}\sum_{p=1}^{P}\sum_{f=1}^{F}\frac{\|\mathbf{s}_p-\mathbf{s}_p^{truth}\|}{\|\mathbf{s}_p^{truth}\|}$, was used instead. The results are shown in Figure 1. From the Figure 1, the relative error of the proposed algorithm is significantly lower than [13] factorization algorithm. The variance of the error is also small, showing that the method is more stable and robust than the original factorization algorithm.

5.2 Qualitative Evaluation on Facial Expressions

Recognizing facial expressions is one of the current challenging problems. Thus, we are motivated to evaluate our approach with facial expressions. In this experiment, a 3D face model with four different expressions captured from 3D Facial Expression Database [15] at the State University of New York was used to examine the qualitative performance of our proposed approach. The four expressions are happy, neutral, sad and surprise. First, we manually selected 68 feature points on the 3D models. Then, the 3D models were rotated about x-axis from -10° to -10° in 2° steps, about y-axis from -20° to -20° in 1° steps and about z-axis from -10° to -10° in 2° steps. In each step, we generated an image of the 3D model. Therefore, we have 4961 images for each expression. Some images with different expressions are shown in Figure 2. The ground truth of the 3D feature points of each expression is shown in Figure 3.

In this experiment, four different levels of Gaussian white noise were added to $\mathbf{W}$, with strength levels of 0%, 5% and 10%. Then, $\mathbf{W}$ is partitioned into 41 subsets for the batch algorithm and each subset is applied factorization with metric optimization. The results are showed in Figure 4, Figure 5 and Figure 6, respectively.

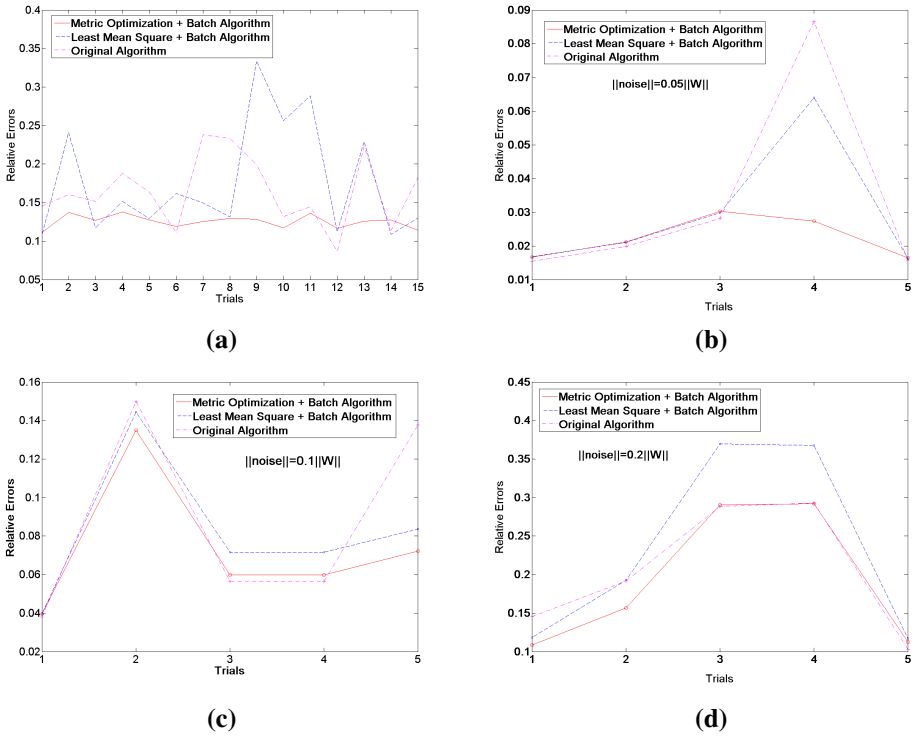


Figure 1: Relative errors of the three different appoaches of the factorization algorithms on rigid synthetic data (*a*) and non-rigid data under different levels of Gaussian white noise (*b*, *c* and *d*).

6 Discussion and Conclusion

Our approach is an extension of the non-rigid factorization algorithm proposed by Xiao et al. [13]. In this paper, we proposed a batch algorithm which uses partitions of the measurement matrix and a metric optimization that recovers the optimized 3D structures based on nonlinear shape constraints. The batch algorithm allows the system to process the data in parallel because the factorization algorithm can be applied on each submatrix separately. Thus, it is suitable for real-time applications such as surveillance and biometric authentication systems. The algorithm does not require to repeatedly compute the factorization algorithm with the whole measurement matrix every time the new data are added. The computation becomes more effective by using our proposed approach.

The metric optimization is another significant contribution in this paper. We introduced the metric optimization to refine the recovered 3D structures by using the new shape constraints. The experiments showed that our approach is more accurate and robust than the existing factorization algorithm for both rigid and non-rigid objects under different strength levels of Gaussian white noise.

References

[1] Matthew Brand. Morphable 3d models from video. *In Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2:456–463, 2001.

[2] Matthew Brand. A direct method for 3d factorization of nonrigid motion observed in 2d. *In Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2:122–128, 2005.

[3] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. *In Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2:690–696, 2000.

[4] C.G. Broyden. The convergence of a class of double-rank minimization algorithm. *Journal Inst. Math. Applic.*, 6:76–90, 1970.

[5] R. Fletcher. A new approach to variable metric algorithms. *Computer Journal*, 13:317–322, 1970.

[6] D. Goldfarb. A family of variable metric updates derived by variational means. *Mathematics of Computing*, 24:23–26, 1970.

[7] Mei Han and Takeo Kanade. Reconstruction of a scene with multiple linearly moving objects. *International Journal of Computer Vision*, 59(3):285–300, 2004.

[8] Conrad J. Poelman and Takeo Kanade. A paraperspective factorization method for shape and motion recovery. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 19(3):206–218, March 1997.

[9] D.F. Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of Computing*, 24:647–656, 1970.

[10] Richard Szeliski and Sing Bing Kang. Recovering 3d shape and motion from image streams using non-linear least squares. Technical Report Series CRL 93/3, Cambridge Research Lab, March 1993.

[11] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[12] L. Torresani, A. Hertzmann, and C. Bregler. Learning non-rigid 3d shape from 2d motion. *In Proc. NIPS 2003*.

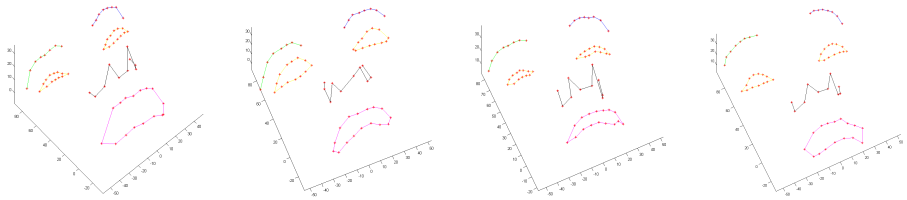
[13] Jing Xiao, JinXiang Chai, and Takeo Kanade. A closed-form solution to non-rigid shape and motion recovery. *International Journal of Computer Vision*, 67(2):233–246, 2006.

[14] Jingyu Yan and Marc Pollefeys. A factorization-based approach to articulated motion recovery. *In Proc. Int. Conf. Computer Vision and Pattern Recognition*, 2:815–821, 2005.

[15] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew Rosato. A 3d facial expression database for facial behavior research. *In Proc. 7th International Conference on Automatic Face and Gesture Recognition*, pages p211–216, 2006.

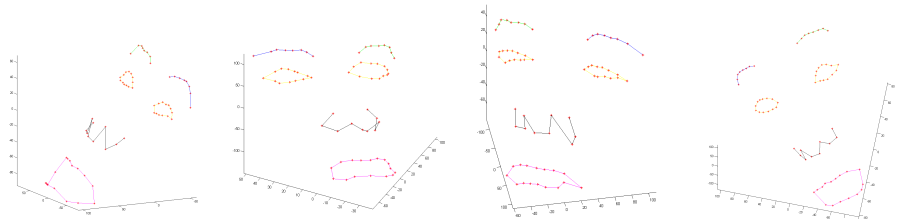
(a) **(b)** **(c)** **(d)**

Figure 2: (a) Happy expression image with rotation about $x = -10^\circ$, $y = 20^\circ$ and $z = 10^\circ$ (b) Neutral expression image with rotation about $x = 0^\circ$, $y = 0^\circ$ and $z = 0^\circ$ (c) Sad expression image with rotation about $x = -10^\circ$, $y = -20^\circ$ and $z = -10^\circ$ (d) Surprise expression image with rotation about $x = -10^\circ$, $y = 20^\circ$ and $z = 10^\circ$.



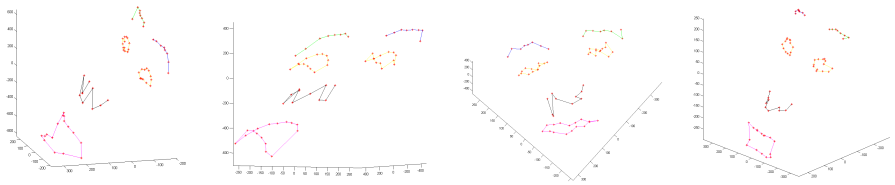
(a) **(b)** **(c)** **(d)**

Figure 3: (a) Ground truth of 3D happy expression (b) Ground truth of 3D neutral expression (c) Ground truth of 3D sad expression (d) Ground truth of 3D surprise expression.



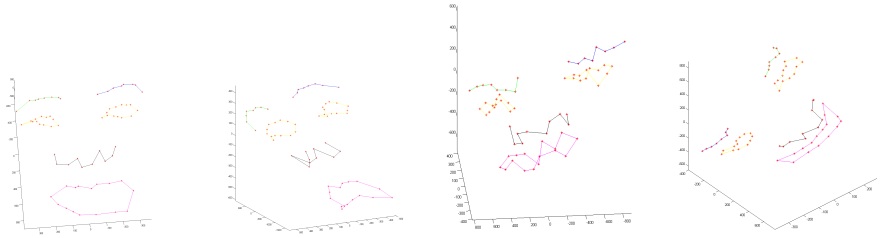
(a) **(b)** **(c)** **(d)**

Figure 4: (a) Reconstructed 3D happy expression (b) Reconstructed 3D neutral expression (c) Reconstructed 3D sad expression (d) Reconstructed 3D surprise expression under 0% Gaussian white noise.



(a) **(b)** **(c)** **(d)**

Figure 5: (a) Reconstructed 3D happy expression (b) Reconstructed 3D neutral expression (c) Reconstructed 3D sad expression (d) Reconstructed 3D surprise expression under 5% Gaussian white noise.



(a) **(b)** **(c)** **(d)**

Figure 6: (a) Reconstructed 3D happy expression (b) Reconstructed 3D neutral expression (c) Reconstructed 3D sad expression (d) Reconstructed 3D surprise expression under 10% Gaussian white noise.