Hand Modeling and Tracking for Video-Based Sign Language Recognition by Robust Principal Component Analysis

Wei Du* and Justus Piater

University of Liège, Department of Electrical Engineering and Computer Science
Montefiore Institute, B28, B-4000 Liege, Belgium
{wei.du,justus.piater}@ulg.ac.be

Abstract. Hand modeling and tracking are essential in video-based sign language recognition. The high reformability and the large number of degrees of freedom of hands render the problem difficult. To tackle these challenges, a novel approach based on robust principal component analysis (PCA) is proposed. The robust PCA incorporates an L_1 norm objective function to deal with background clutter, and a projection pursuit strategy to deal with the lack of alignment due to the deformation of hands. The learning algorithm of the robust PCA is very simple, involving only a search for the solutions in a finite set constructed from the training data, which leads to the learning of much more representative and interpretable bases. The incorporation of the L_1 regularization in the fitting of the learned robust PCA models results in cleaner reconstructions and more stable fitting. Based on the robust PCA, a hand tracking system is developed that contains a skin-color region segmentation based on graph cuts and template matching in the framework of particle filtering. Experiments on a publicly available sign-language video database demonstrates the strength of the method.

Keywords: hand modeling and tracking, sign language recognition, robust PCA, L_1 norm.

1 Introduction

Automated sign language recognition from video has been studied for at least about twenty years [1]. However, the recognition of continuous, natural signing remains challenging, in terms of video analysis, due to the multimodal nature of the cues including hands, lips, facial expressions and body poses.

In this paper, we concentrate on one particular challenge, hand modeling and tracking, in video-based sign language recognition, as hands convey a lot of information including at least configurations, positions, and instantaneous

* The research in this paper has received funding from the European Community's Seventh Framework Programme FP7/2007–2013 – Challenge 2 - Cognitive Systems, Interaction, Robotics – under grant agreement n° 231424-SignSpeak.

K.N. Kutulakos (Ed.): ECCV 2010 Workshops, Part I, LNCS 6553, pp. 273–285, 2012.

velocities. In principle, hands are difficult to model and to track because of their high deformability, their large number of degrees of freedom and their high level of self-occlusion, which give rise to an enormous variation of appearance and a high level of ambiguity.

Previous approaches have tackled these difficulties by different methodologies including exploiting the skin color [2,3], discriminative learning of hand classifiers [4,5] and incorporating constraints from arms [6] and from TV subtitles [7,8], etc. For hand modeling and recognition, principal component analysis (PCA) has also been applied and achieved promising results [9,10]. However, clean hand segmentation and clean backgrounds are needed as classical PCA is known to be sensitive to outliers.

This paper presents an approach to modeling and tracking hands in sign language videos using PCA. In contrast to previous work, we directly deal with the outliers introduced by background clutter and by the deformation of hands. To this end, we propose a novel, robust PCA that incorporates a robust objective function based on the L_1 norm and limits the search space to a finite set constructed only from training data, which results in the learning of much more representative and interpretable PCA bases. An efficient algorithm is proposed based on iterative projection pursuit to solve the robust PCA learning problem. We also incorporate the L_1 regularization in the fitting of the learned, robust PCA models, leading to cleaner reconstructions and more stable fitting. Based on the robust PCA, we develop a hand tracking system that contains a skin-color region segmentation based on graph cuts and template matching in the framework of particle filtering. Experiments on a publicly available sign-language video database demonstrate the strength of the method.

We introduce our robust PCA algorithm in Section 2. Section 3 presents how to apply the robust PCA method for hand modeling and tracking. Experimental results are shown in Section 4. Section 5 gives the conclusions and discusses possible future work.

2 Robust Principal Component Analysis

Principal component analysis (PCA) is one of the most popular tools for high-dimensional data analysis where dimensionality reduction is necessary to reduce the number of input variables in order to simplify the problems. Commonly, in PCA, one tries to find a set of projections that maximize the variance of given data, or equivalently, that minimize the residuals of the projections. These projections constitute a low-dimensional, linear subspace in which the data structure in the original input space can effectively be captured.

However, although successful in many applications such as face recognition [11], classical PCA is known to be sensitive to outliers [12,13,14,15], since the computation of both data variances and projection residuals is founded on the L_2 norm which exaggerates the effects of outliers with a large norm. To achieve robustness, many approaches have been proposed, including replacing the L_2 norm by the L_1 norm [12,13] or other robust scalar estimators [14], and adding

a robustfying term in analogy to regularization in PCA [15]. Although these different PCA models optimize different objective functions, the optimization is done over all training data so that the learned bases can be considered as some kind of combination of the data.

In the problem of sign language recognition, a fundamental difficulty in extending these concepts to modeling and tracking hands is that, unlike faces, hands are highly deformable and thus are difficult to align. A consequence is that the hand training examples cropped from sign language videos contain not only hands in various configurations but also some background clutter. In the presence of missing alignments in the training data, the PCA bases learned by most methods will be blurred and carry no representative information to model the objects of interest.

Here, we propose a novel PCA method that combines a few strategies to deal with the alignment difficulties. Similar to previous work, we also search the PCA bases that maximize the spread of the training data. However, we consider the background clutter as outliers and incorporate a robust objective function based on the L_1 norm. Then, instead of searching the entire space of possible directions for the bases, we only check for vectors belonging to a finite set constructed from training data. The intuition behind this is that if the training set is large enough, there is good hope that quite some of the data will be close to the directions of maximal spread. After the PCA modeling, we incorporate the L_1 regularization into the fitting of the PCA models, which greatly improves the fitting quality.

2.1 Problem Formulation

Let $X_{d\times n} = \{x_1, \ldots, x_n\}$ be the training data matrix with d the dimension of the data and n the number of training samples. Assume that the data has been centered by removing the mean. In PCA, the bases to be learned are those orthogonal directions in which the training data are the most widely spread. Classical PCA defines the spread by the variance of the projected data, which is sensitive to outliers. We incorporate a more robust definition of spread [13] that is based on the L_1 norm, given by

$$S(X, w) = ||w^T X||_1 = \sum_{i=1}^{n} |w^T x_i|, \tag{1}$$

where w is a unit vector representing the projection direction.

Thus, our robust PCA amounts to optimizing

$$W = \text{argmax}_W ||W^T X||_1 = \text{argmax}_W \sum_{j=1}^{k} \sum_{i=1}^{n} |w_j^T x_i|, \text{ s.t. } W^T W = I, \tag{2}$$

where $W_{d\times k} = \{w_1, \ldots, x_k\}$ is the PCA basis matrix, with k being the total number of bases.

Input: X, k
X: training data set
k: number of bases
Output: $W = \{w_1, \ldots, w_k\}$
Compute the mean of the data $\hat{u}(X)$
Set $x_i^1 = x_i - \hat{u}(X),\ i = 1, \ldots, n$, and $A^1 = \left\{\frac{x_i^1}{||x_i^1||}, i = 1, \ldots, n\right\}$.
Compute the first PCA basis

$$w_1 = \text{argmax}_{w_1 \in A^1} \sum_{i=1}^{n} |w_1^T x_i|$$

and compute the projections on the first basis as $y_i^1 = w_1^T x_i^1,\ i = 1, \ldots, n$.
for $l = 2 : k$ **do**
  Set $x_i^l = x_i^{l-1} - y_i^{l-1} w_{l-1},\ i = 1, \ldots, n$ and $A^l = \left\{\frac{x_i^l}{||x_i^l||}, i = 1, \ldots, n\right\}$.
  Compute the lth PCA basis

$$w_l = \text{argmax}_{w_l \in A^l} \sum_{i=1}^{n} |w_l^T x_i|$$

  and compute the projections as $y_i^l = w_l^T x_i^l,\ i = 1, \ldots, n$.
end

Algorithm 1. Iterative Projection Pursuit for Robust PCA.

2.2 Optimization Algorithm

In practice, eq. 2 is difficult to optimize due to the non-differentiability of the L_1 norm and the orthogonal constraint in the bases. In [13], a simple iterative algorithm is proposed to solve the L_1-norm maximization in eq. 2. However, as mentioned earlier, the optimization over all training data leads to a solution in the form of some combination of the data which is unsuitable when the data are not aligned.

To deal with the lack of alignment, we adopt a projection pursuit strategy [14][1] that, instead of searching in the whole space of possible directions for the bases, iteratively selects the bases from a candidate set constructed from the training data. See algorithm 1 for details.

The data mean $\hat{u}(X)$ in the algorithm is estimated by the spatial median or L_1-median, defined as

$$\hat{u}(X) = \text{argmin}_u \sum_{i=1}^{n} ||x_i - u||. \tag{3}$$

[1] Note that this work [14] was originally motivated only by outliers instead of the alignment problem. The lack of alignment makes some inliers look like outliers. Another difference is that in this work [14], the objective function to optimize is the median absolute deviation (MAD) instead of the L_1 norm here. MAD has a higher number of breakdown points and is thus more robust, but is more difficult to optimize than the L_1 norm.

Put in words, the algorithm iteratively projects the data into the subspace of the orthogonal complement to the existing bases and then finds the most representative among the normalized projected data that maximizes the L_1 norm. In order for this algorithm to work, the candidate set A^l in which the bases are searched for should be quite dense in the region where the L_1 norm reaches its maximum, which holds true if there are enough training samples such that some of them are close to the direction where data is widely spread.

Although the above optimization algorithm is extremely simple, its complexity is quadratic in the size of the candidate set A^l, or equivalently, in the number of the training samples. As mentioned above, the algorithm requires a relatively large number of training samples to work, and is thus not suitable for online learning where fast and incremental training is necessary, or in cases where training data are sparse.

2.3 Fitting with the Robust PCA Models

After the learning of the robust PCA bases, the fitting of a new given sample x^* is simply the projection of the data onto the learned bases. However, we found that incorporating the L_1 regularization into the fitting leads to sparse reconstructions that include only a small number of the linear bases, which is helpful in dealing with the alignment problem. In particular, we solve the following L_1-regularized L_2 fitting loss function

$$y^* = \text{argmin}_y ||y^T W - x^*||_2 + \lambda_2 ||y||_2 + \lambda_1 ||y||_1, \tag{4}$$

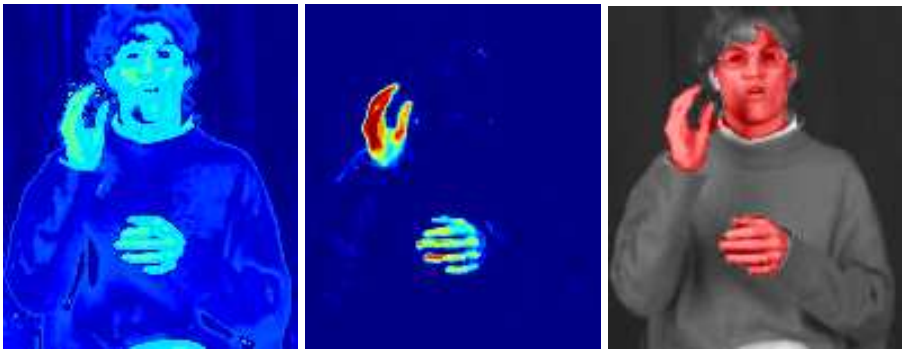
where λ_2 and λ_1 are the L_2 and L_1 regularization coefficients. Eq. 4 is also called an *elastic net* [16] that balances the smoothness and the sparsity of the solutions, which can be efficiently solved by using e.g. interior-point methods [17]. Note that $y^{*T} W$ gives the reconstruction of x^*.

3 Hand Modeling and Tracking by Robust PCA

The application of the above robust PCA for hand modeling in sign language recognition is straightforward, provided that a large number of training hand samples are available. Below, we will explain how the learned PCA models can be used for hand tracking. In short, our hand tracking system contains two steps including skin-color region segmentation followed by PCA-based template matching in a particle-filtering framework.

3.1 Skin-Color Region Segmentation

For the segmentation of the skin regions, the popular graph-cut algorithm is adopted [18]. Graph cuts seek to minimize an energy function of the form



Color likelihood Motion likelihood Segmentation

Fig. 1. Color- and motion-based face and hand segmentation

$$E = \sum_{p \in P} D_p(x_p) + \sum_{\{p,q\} \in N} V_{p,q}(x_p, x_q),$$

where D_p is called the data or unary term that measures how well label x_p fits pixel p given the observed data, and $V_{p,q}$ is called the smoothness or pairwise term that enforces smooth labeling among neighboring pixels.

For our skin segmentation problem, we incorporate two types of information in D_p. The first is a color likelihood based on histogram matching, and the second is a motion likelihood based on image differencing. The intuition behind this is that hands of signers have distinct skin colors that are different from the background, and that the hands produce the most dramatic movement in sign language videos (Fig. 1). For the smoothness term, we adopt the contrast-sensitive Potts model [19],

$$V_{p,q}(x_p, x_q) = \begin{cases} 0 & \text{if } x_p = x_q \\ \alpha + \beta \exp(-\frac{||I_p - I_q||^2}{\theta}) & \text{otherwise} \end{cases},$$

where I_p and I_q are the color vectors of pixels p and q respectively, and α, β, θ are model parameters whose values are learned using training data. One example of skin segmentation is illustrated in Fig. 1.

3.2 PCA-Based Template Matching and Particle Filtering

After segmentation, we search hands in only the segmented skin regions using PCA-based template matching in a framework of particle filtering [20]. We first sample a number of hand candidates from skin regions, and match them with the PCA bases of the left and right hands. Thus, two matching scores are computed for each hand candidate reflecting the probability that the candidate is the left and the right hand. The hand model with the highest match score is most likely to be the hand being tracked in the current frame. However, we smooth hand trajectories over time by penalizing large motions between frames. This is currently done offline using dynamic-programming techniques [21].

man 1 woman 1 woman 2

Fig. 2. RWTH-Boston-104 database. This database is publicly available at `http://www-i6.informatik.rwth-aachen.de/aslr/database-rwth-boston-104.php`.

4 Experimental Results

We demonstrate the effectiveness of our robust PCA based hand modeling and tracking on the RWTH-Boston-104 database [22]. The database contains 201 short sign language video sequences of about 100 frames taken from three signers under a controlled situation. The manually-labeled hand positions of the entire database are available and are used to collect training hand samples. Figure 2 shows three example images from some sequences of the three signers. Note that the male signer sits closer to the camera and some parts of the hands are cropped and occluded by the annotation bars at the bottoms of the images.

4.1 Results of Hand Modeling by Robust PCA

We first tested our robust PCA for hand modeling. Twenty videos from each signer were selected and used to learn the signer-specific PCA models. The training hand samples were collected by cropping a subwindow of size 51 by 51 around the hand positions in each frame of a training sequence. Overall, 100 bases are extracted for the left and the right hands of each signer. Figure 3 shows some examples of the training data for the male signer. Note that the two hands are occluded by the annotation bar in the beginning and in the end of the sequence and that these occluded samples are used as well in training the PCA models, making the problem of outliers worse.

Figures 4 and 5 show the learned robust PCA models and the corresponding models learned by classical PCA for the left and the right hands of the male signer. The differences between them are quite clear. As classical PCA and many other robust variations learn the bases as a function of all training data, the lack of alignment in the data causes the bases to be blurred and less representative. In contrast, our robust PCA searches the bases by selecting among the data the most representative direction in each iteration. Thus, the learned bases are much more meaningful and interpretable. The robust PCA models learned for the two female signers are shown in Figure 6 and 7. Figure 8 shows the reconstructions of a test sequence by using the learned PCA models for the male signer. Recall that the reconstruction is done by solving eq. 4.

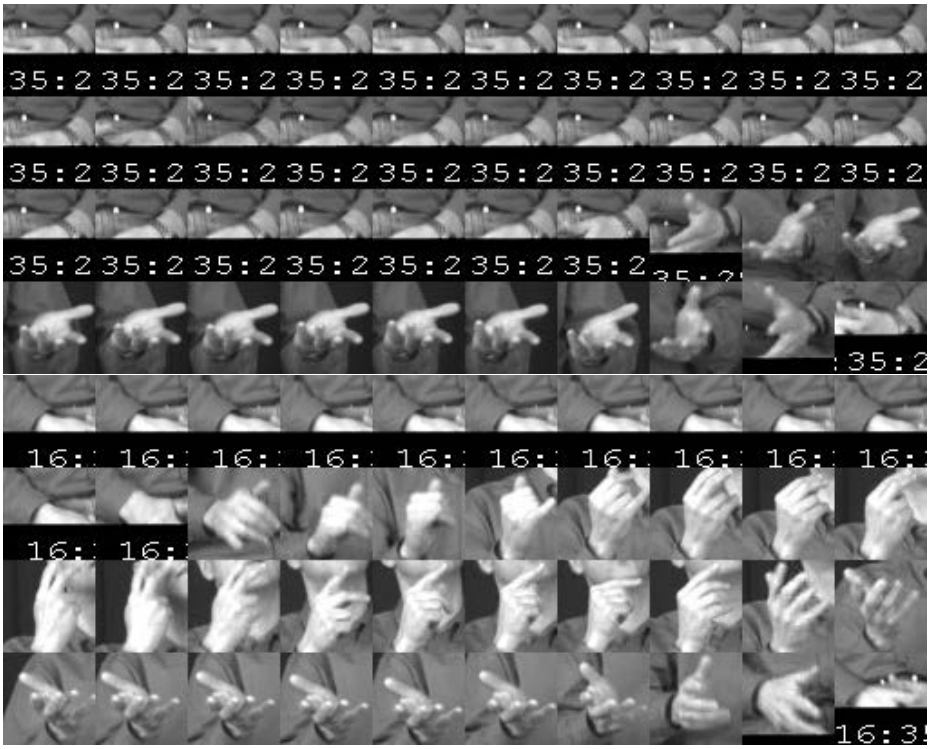


Fig. 3. Some training hand examples extracted from a sequence of the male signer. Note that the occlusions from the annotation bars are stronger for the left hand than for the right hand, since the left hand moves less often and less dramatically. This also holds true for most other sequences in the database. Hand regions cropped from the entire sequence are used for training the PCA models.

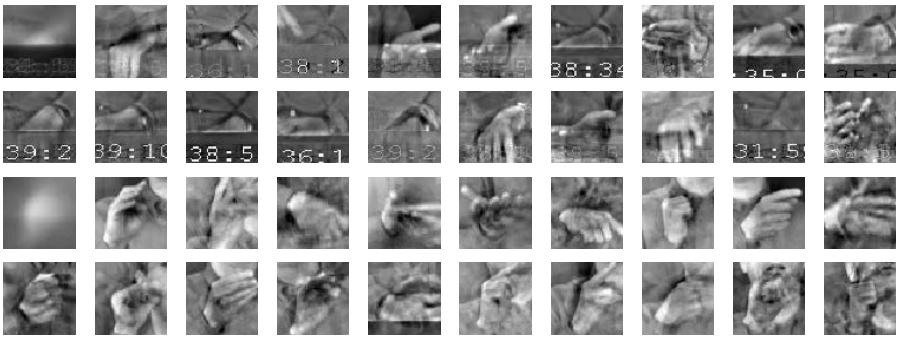


Fig. 4. PCA bases learned by our robust PCA. The first 20 bases of the left and the right hands of the male signer are shown in the top and bottom two rows.

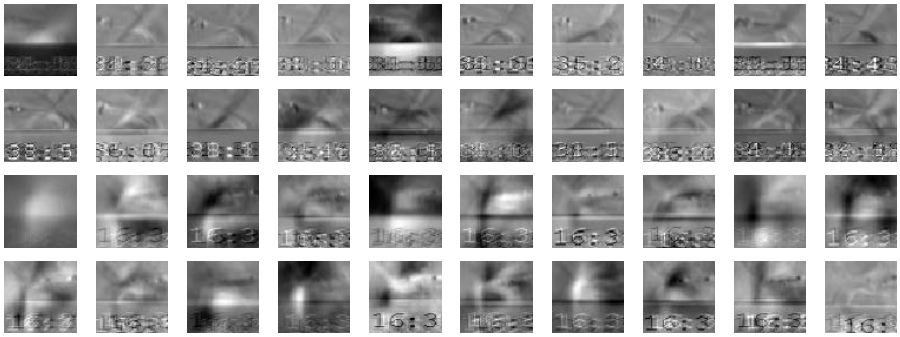


Fig. 5. PCA bases learned by classical PCA. The first 20 bases of the left and the right hands of the male signer are shown in the top and bottom two rows.

Fig. 6. PCA bases learned by our robust PCA. The first 20 bases of the left and the right hands of one female signer are shown in the top and bottom two rows.

Fig. 7. PCA bases learned by our robust PCA. The first 20 bases of the left and the right hands of the other female signer are shown in the top and bottom two rows.

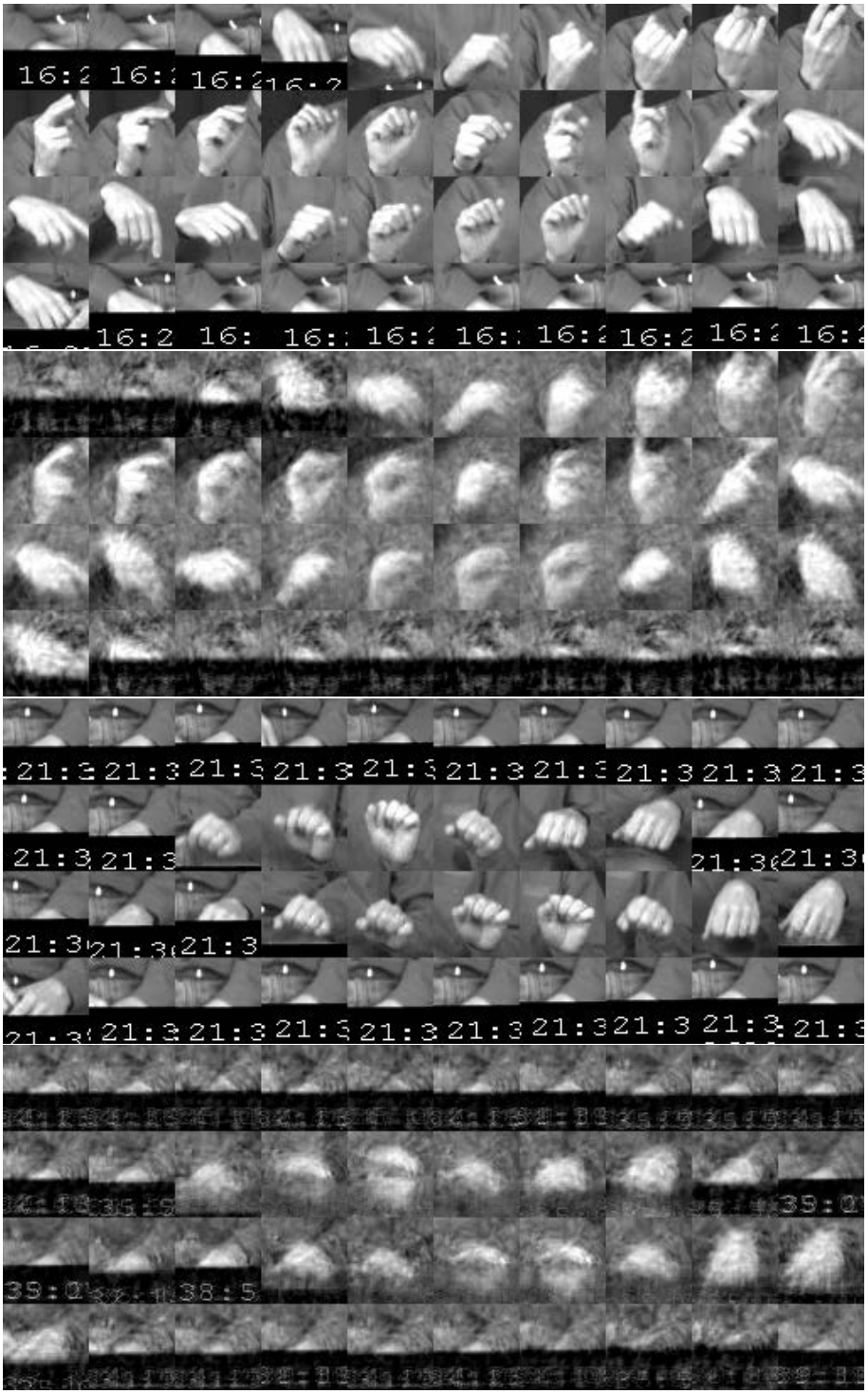


Fig. 8. Some test examples from a sequence and their reconstructions by using the learned robust PCA bases and the L_1 regularized least-squares solver.

4.2 Results of Hand Tracking

As the segmentation of the hands by using graph cuts is quite good, tracking is not a very difficult task for this database. In the framework of particle filtering, 200 samples are generated from segmented hand regions and another 100 samples are generated from motion predictions. The evaluation of the particles are done by computing a fitting score for each particle with the learned PCA hand models by solving eq. 4 with $\lambda_2 = 0.01$ and $\lambda_1 = 0.01$ using an L_1 regularized L_2 least-squares solver [17]. As quite a large number of particles need to be evaluated, tracking is not real time and takes a few minutes to process one video of 100 frames on an ordinary laptop with a dual-core CPU of 2.66G and 4G memory. Some tracked hand regions and the corresponding PCA reconstructions are shown in Fig. 9. Hand trajectories of one video sequence are shown in Fig. 10.

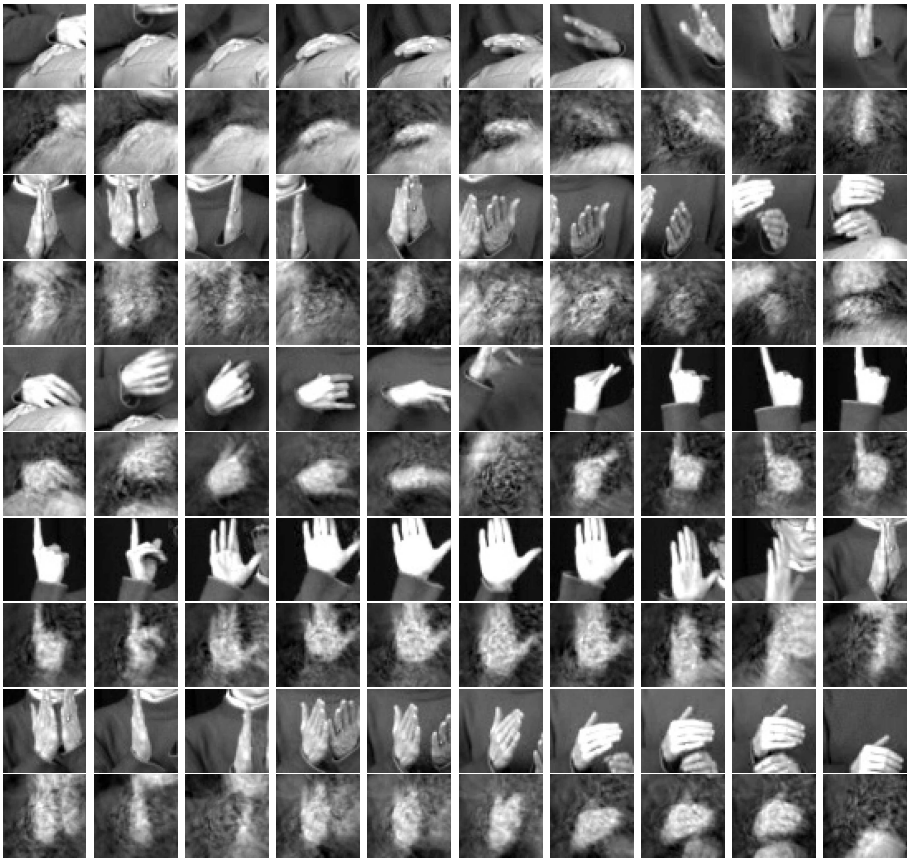


Fig. 9. The tracked hand regions, odd rows, and the PCA reconstructions, even rows

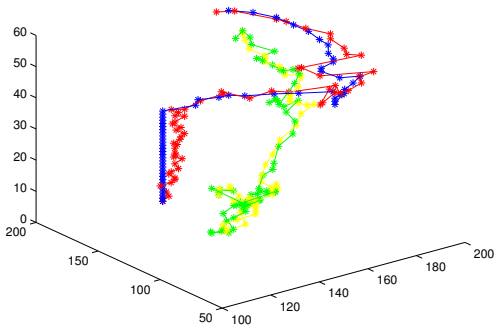


Fig. 10. Hand trajectories in a video sequence of 60 frames. Blue and yellow are the ground truth of the left and right hands, and red and green are the tracking results.

5 Conclusions and Future Work

This paper presents a novel robust PCA for hand modeling and tracking in video-based sign language recognition. The motivation was to handle the high deformability of the hands which renders the alignment of the hand data difficult, if not impossible. The strength of the method, demonstrated on a publicly-available database, roots in the incorporation of the robust L_1 norm and the projection pursuit strategy involving only the searching of the solutions in a finite set constructed from the training data.

Empirically, our robust PCA achieved much more representative and interpretable bases than classical PCA in the presence of outliers and missing alignments. A hand tracking system was developed based on the robust PCA and obtained promising results. However, the tracking system needs to be enhanced. For instance, currently, the left and right hands are tracked separately with no explicit modeling of the interactions between them, which is not a big problem as the sequences in the database are short and the segmentation by graph cuts is generally good. Some switches of the trackers of the two hands were corrected by the incorporation of some heuristics and by the offline processing of dynamic programming. For long-term tracking, the knowledge on the configuration of the upper body would help disambiguate the hands during heavy interactions [6]. The learned robust PCA bases seem to be useful for classifying hand configurations, which is also worth further investigation.

References

1. Dorner, B.: Hand shape identification and tracking for sign language interpretation. In: IJCAI Workshop on Looking at People (1993)
2. Starner, T., Weaver, J., Pentland, A.: Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 1371–1375 (1998)

3. Cooper, H., Bowden, R.: Large Lexicon Detection of Sign Language. In: Lew, M., Sebe, N., Huang, T.S., Bakker, E.M. (eds.) HCI 2007. LNCS, vol. 4796, pp. 88–97. Springer, Heidelberg (2007)
4. Kadir, T., Bowden, R., Ong, E.J., Zisserman, A.: Minimal training, large lexicon, unconstrained sign language recognition. In: British Machine Vision Conference, Kingston, UK (2004)
5. Ong, E., Bowden, R.: A boosted classifier tree for hand shape detection. In: Internatial Conference on Automatic Face and Gesture Recogntion (2004)
6. Buehler, P., Everingham, M., Huttenlocher, D., Zisserman, A.: Long term arm and hand tracking for continuous sign language TV broadcasts. In: British Machine Vision Conference (2008)
7. Cooper, H., Bowden, R.: Learning Signs from Subtitles: A Weakly Supervised Approach to Sign Language Recognition. In: Computer Vision and Pattern Recognition, pp. 2568–2574 (2009)
8. Buehler, P., Everingham, M., Zisserman, A.: Learning sign language by watching TV (using weakly aligned subtitles). In: Computer Vision and Pattern Recognition (2009)
9. Coogan, T., Sutherland, A.: Transformation invariance in hand shape recognition. In: International Conference on Pattern Recognition (2006)
10. Huang, D.Y., Hu, W.C., Chang, S.H.: Vision-based hand gesture recognition using pca+gabor filters and svm. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 1–4 (2009)
11. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)
12. Ding, C., Zhou, D., He, X., Zha, H.: R1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization. In: ICML 2006: Proceedings of the 23rd International Conference on Machine Learning, pp. 281–288 (2006)
13. Kwak, N.: Principal component analysis based on L1-norm maximization. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 1672–1680 (2008)
14. Croux, C., Ruiz-Gazen, A.: High breakdown estimators for principal components: the Projection-pursuit approach revisited. Journal of Multivariate Analysis 95, 206–226 (2005)
15. La Torre, F.D., Black, M.J.: A framework for robust subspace learning. International Journal of Computer Vision 54, 117–142 (2003)
16. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67, 301–320 (2005)
17. Kim, S.J., Koh, K., Lustig, M., Boyd, S., Gorinevsky, D.: An interior-point method for large-scale l1-regularized least squares. IEEE Journal on Selected Topics in Signal Processing 4, 606–617 (2007)
18. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23, 1222–1239 (2001)
19. Boykov, Y., Jolly, M.: Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In: International Conference on Computer Vision, vol. I, pp. 105–112 (2001)
20. Doucet, A., de Freitas, N., Gordon, N.: Sequential Monte Carlo Methods in Practice. Springer, New York (2001)
21. Godsill, S., Doucet, A., West, M.: Maximum a posteriori sequence estimation using Monte Carlo particle filters. Annals of the Institute of Statistical Mathematics 53, 82–96 (2001)
22. Dreuw, P., Rybach, D., Deselaers, T., Zahedi, M., Ney, H.: Speech Recognition Techniques for a Sign Language Recognition System. In: Interspeech, pp. 2513–2516 (2007)