Dimensionality reduction for efficient single frame hand pose estimation

Petros Douvantzis, Iason Oikonomidis, Nikolaos Kyriazis, and Antonis Argyros

Institute of Computer Scince - FORTH,
Computer Science Department - University of Crete
Heraklion, Crete, Greece

Abstract. Model based approaches for the recovery of the 3D position, orientation and full articulation of the human hand have a number of attractive properties. One bottleneck towards their practical exploitation is their computational cost. To a large extent, this is determined by the large dimensionality of the problem to be solved. In this work we exploit the fact that the parametric joints space representing hand configurations is highly redundant. Thus, we employ Principal Component Analysis (PCA) to learn a lower dimensional space that describes compactly and effectively the human hand articulation. The reduced dimensionality of the resulting space leads to a simpler optimization problem, so model-based approaches require less computational effort to solve it. Experiments demonstrate that the proposed approach achieves better accuracy in hand pose recovery compared to a state of the art baseline method using only 1/4 of the latter's computational budget.

Keywords: Model based hand pose estimation, dimensionality reduction, PCA

1 Introduction

The problem of estimating the configuration of the human body or its parts has high practical and theoretical interest. Body pose can be calculated at several scales and granularities including full body [14, 18], head [13] and hand pose [3, 4, 10, 12, 19]. Recovering the articulation of a human hand can be proven very useful in a number of application domains including but not limited to advanced HCI/HRI, games, AR applications, sign language understanding, etc. Depending on the application requirements, different variants of the basic hand pose[1] estimation problem can be formulated. If a sequence of poses needs to be estimated, the resulting problem can be described as *hand pose tracking*. If a single hand pose needs to be estimated from a single frame, without prior knowledge resulting from temporal continuity, the problem is referred to as *single frame hand pose estimation*. Clearly, a solution to the single frame hand pose

[1] In this work, we use the term *hand pose* to refer to the 3D position, orientation and full articulation of a human hand.

estimation problem can be used to bootstrap hand pose tracking. Additionally, hand pose tracking can be implemented as sequential hand pose estimation.

Several methods have been proposed to solve the hand pose estimation problem using markerless visual data. According to [3] the existing approaches can be categorized as partial or full pose estimation methods, depending on the completeness of the output. The last class is further divided to *appearance-based* and *model-based* methods. Appearance-based methods [15, 16, 18, 20] estimate hand configurations directly from images, using a precomputed mapping from the image feature space to the hand configuration. Model-based methods [2, 4, 5, 10–12] search the solution space of the problem for the hand configuration that is most compatible to the observed hand. Every hypothesized hand pose is transferred to feature space using a hand shape model that is compared to the observed features.

Appearance-based methods are computationally efficient at estimation phase and usually do not require initialization. However, their accuracy is proportional to the wealth of the training set. Model-based approaches have better resolution since they search the parameter space without discretizing it. Their primary limitation is their computational requirements which are due to the high dimensionality of the problem to be solved. A large number of hypothesized poses must be compared to the observed data during the search of the problem space. For the case of hand pose tracking, due to temporal continuity, the current solution needs to be searched in the vicinity of the solution of the previous frame. This has resulted in methods that achieve near real time performance [10]. However, in single frame hand pose estimation, the position, orientation and articulation of the human hand needs to be estimated without a prior knowledge of the observed hand's configuration and, thus, the full parametric space needs to be searched.

In this work, we come up with a compact representation of hand articulation that permits model based methods to operate on lower dimensional spaces and become more efficient. Adopting the representation used in [10], the articulation of the human hand can be described by 20 parameters encoding the angles of the human hand joints and thus forming a 20-dimensional space. However, only a fraction of this space contains valid hand configurations [6, 17]. For example, if no external forces are applied, the range of motions that can be performed by each finger joint is limited [9]. It may also be the case that a combination of plausible joint values may result in implausible hand configurations. As an example, a combination of joint values may result in a finger intersecting itself, another finger or the palm. Furthermore, the biomechanics and the physiology of the hand impose inter-finger and intra-finger constraints. For example, bending the pinky causes the ring finger to be bent as well. Last but not least, when a human hand is known to be engaged in specific activities (e.g., grasping, sign language, etc), its configurations are expected to lie in a much lower dimensional manifold.

Most of the mentioned constraints are hard to model analytically. However, they can be learnt by employing dimensionality reduction techniques on sets of

training samples. The reduced spaces model implicitly the existing constraints. In this work we perform dimensionality reduction by employing Principal Components Analysis (PCA). By providing mechanisms to move between representations, model-based methods can operate on the transformed space of reduced dimensionality. Although dimensionality reduction and PCA has been employed in the past for hand pose estimation [1, 7], in this work we evaluate the performance gain obtained for the problem of single frame hand pose estimation and demonstrate its integration with an existing, state of the art baseline method. The model-based method employed as the baseline is that of Oikonomidis et al. [10]. For hand pose tracking, this method achieves an accuracy of $5mm$ and a computational performance of 20 fps. However, due to the dimensionality of the problem, for single frame hand pose estimation the method is more time consuming and less reliable. For this reason, tracking is performed only after the hand is manually initialized for the first frame. In our work we show that by employing dimensionality reduction, the problem's search space is reduced to such a degree that single frame hand pose estimation becomes practical. More specifically, experimental results demonstrate that the proposed approach estimates the hand pose accurately and only at a fraction of the computational budget required by the baseline method. Thus, the need for manual initialization of the pose of the hand is lifted and the potential of model based methods to support hand pose estimation to real-world applications in considerably increased.

2 Proposed Approach

According to the employed baseline method [10], hand pose estimation is formulated as an optimization problem, seeking for the hand model parameters that minimize the discrepancy between the appearance of hypothesized hand configurations and the actual hand observation. Observations come from images acquired by an RGB-D camera and consist of estimated depth and skin color maps. Hypotheses are rendered through graphics techniques that also give rise to features comparable to those of the observations. Hypotheses are evaluated with an objective function that measures their compatibility to the observations. The optimization problem is handled by a variant of Particle Swarm Optimization (PSO) [8] (optimization module), which searches the parametric space of hand configurations.

A hand pose is represented by a 27-dimensional vector. 3 parameters encode the 3D position of the hand, 4 parameters its 3D orientation in quaternion representation and 20 parameters encode its articulation. Thus, the baseline method searches the 27-dimensional space S_{27}.

The proposed method incorporates PCA in the baseline method and creates a reduced dimensionality search space S_{7+M} consisting of $7 + M$ dimensions. 7 parameters encode 3D position and 3D orientation, while M parameters encode the articulation in a M-dimensional PCA space ($M \leq 20$). Two variations are proposed for the problem of single frame hand pose estimation, namely single-PCA and multi-PCA.

2.1 Single-PCA

Single-PCA consists of a training and an estimation phase. At the training phase of the algorithm, a single PCA space of M dimensions is created as follows. A dataset $\boldsymbol{A}$ consisting of N hand configurations $\boldsymbol{h}$ of 20 dimensions each is used as a training set. After the data are standardised, the orthogonal basis $\boldsymbol{W}$ of the PCA space is calculated. The orthogonal basis $\boldsymbol{W}$, along with the mean and standard deviation per dimension contain all the information needed for projection and back-projection to/from the PCA space. From now on, these elements are considered to form a trained PCA model P.

The number of dimensions M of P that are sufficient to describe the articulation information of the training set can be decided by calculating the cumulative variance explained by the first M PCA dimensions. Otherwise, the number of dimensions M can be chosen experimentally. A small number M introduces a representation error in the PCA space, but results in a smaller space to be searched.

Having encoded the articulation in a M dimensional space, we do not know the range of values of the newly defined PCA dimensions. This information is important because PSO requires the knowledge of the ranges of parameter values during optimization. Hence, we calculate the lower and upper bounds per PCA dimension as follows. Using the trained PCA model P, the $N \times 20$ matrix $\boldsymbol{A_{20}}$ is projected to a $N \times M$ matrix A_M. The standard deviation σ per PCA dimension of A_M is calculated and the lower/upper bounds are estimated as

$$[\boldsymbol{b}_L, \boldsymbol{b}_H] = [-2\boldsymbol{\sigma}, +2\boldsymbol{\sigma}] \ . \tag{1}$$

During the estimation phase, the baseline method is modified as shown in Fig. 1. The optimizer searches the bounded space S_{7+M} using the hypothesis evaluation module to calculate the objective function for the given observation. It should be noted that since the hypothesis evaluation module (see Fig.1) evaluates 27-dimensional parameter hypotheses, a hypothesis in space S_{7+M} must first be back-projected to the original space S_{27}. Finally, the best scoring hypothesis across all PSO generations is back-projected to S_{27} and returned as the solution.

2.2 Multi-PCA

PCA is not sufficient to represent non linearly correlated data such as multiple hand configurations. However, since most hand motions reside in some linear subspace [3], a different PCA model can be trained for each of them. This idea is used in multi-PCA.

More specifically, during the training phase, we use multiple articulation training sets $\boldsymbol{A^i}$, $i = 1, 2, .., F$. Each set consists of data that are linearly correlated. A different PCA model P^i is trained for each set $\boldsymbol{A^i}$, producing F bounded spaces S^i to be searched by the optimizer. Thus, during the estimation phase, F single-PCA optimization problems are solved independently. The hypothesis with the lowest error across all F solutions is chosen as the final solution.

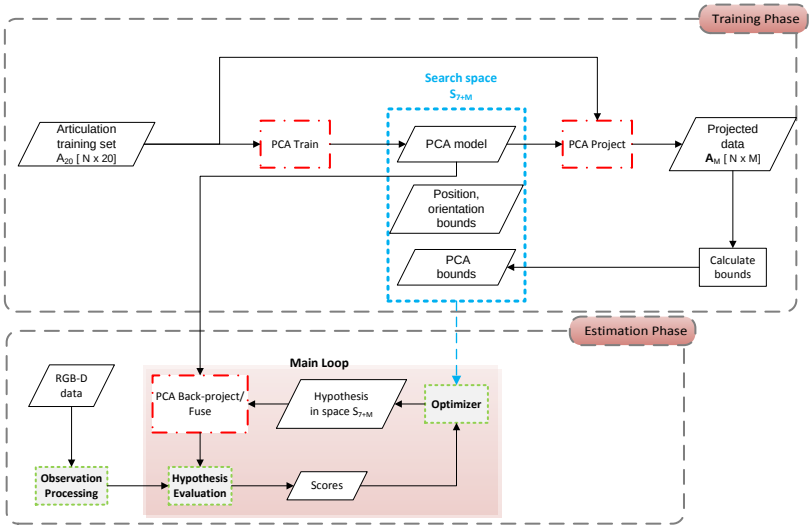


Fig. 1: Graphical illustration of the proposed methodology. At training phase, the reduced dimensionality search space S_{7+M} is created. At estimation phase, an RGB-D frame is used as input and the best hypothesis is searched for. The enclosed modules at the bottom, form the algorithm's main loop which runs for each generation of the optimizer (PSO).

3 Experimental Evaluation

The quantitative evaluation of single-PCA and multi-PCA is based on synthetic data, which enable the assessment of the proposed method against known ground truth. In this direction, a test set consisting of 27-dimensional configurations was rendered as RGB-D images, simulating a real sensor acquiring hand observations.

To quantify the accuracy in hand pose estimation, we employ the metric adopted in [10]. More specifically, the hand pose estimation error Δ is the averaged Euclidean distance between the respective phalanx endpoints of the ground truth and the estimated hand pose.

All the experiments ran offline on a computer equipped with a quad-core intel i7 930 CPU, 6 GBs RAM and an Nvidia GTX 580 GPU with 1581 GFlops processing power and 1.5 GBs memory.

3.1 Dataset creation

The single-PCA algorithm requires a single training set, while the multi-PCA algorithm requires multiple. In our experiments, single-PCA was trained on the union of the training sets of multi-PCA. More specifically, we created three sequences of 100 hand poses each corresponding to one of the following three different hand motions: pinching, open hand closing towards a cylindrical grasp, open

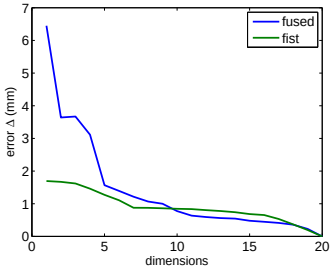


Fig. 2: The representation error of PCA with respect to the number of dimensions for the fused dataset and the fist dataset.

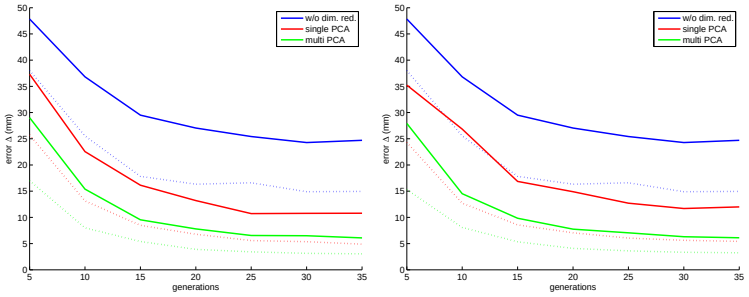
hand closing towards a closed fist. For each of the datasets, a training set was created by adding uniform noise of 0.1 rad in each dimension. Multi-PCA used each training set separately, while single-PCA used them as a unified training set. The test set was created from the original three sequences by adding noise as before and by providing a position and orientation for each pose. The final test set contained 75 poses, uniformly sampled from each of the three sequences.

Since a representation error is introduced due to dimensionality reduction, it is useful to examine the PCA's behaviour when applied to the employed dataset. This enables us to roughly estimate the proposed method's accuracy in single frame hand pose estimation, since the representation error can be considered as a lower limit to the error achieved by single-PCA and multi-PCA. We are interested in the behaviour of PCA on the three separate datasets (multi-PCA case) and on the fused dataset (single-PCA case). We expect the latter to perform worse due to its more complicated contents, because more complex data need more PCA dimensions to account for their variance.

Towards this direction, PCA models are learned for each training set with varying numbers of PCA dimensions $M = 1, 2, .., 20$. In order to measure the error introduced by the reduced dimensionality, each data sample $\boldsymbol{h}$ of the test set is projected to the PCA space S_{7+M} and then back-projected to a point $\boldsymbol{h^{bp}}$ onto the original space S_{27}. The error $\Delta(h, h^{bp})$ is calculated for all samples. The averaged error is shown in Fig. 2. More dimensions are required by the fused dataset to achieve as low representation error as the fist dataset. Thus, we can expect multi-PCA to require a smaller number of dimensions than single-PCA for the same estimation accuracy.

3.2 Results

We evaluated the single-PCA and the multi-PCA method comparatively to the baseline method. The algorithms ran three times for each frame of the test set. The algorithms' behaviour is examined with respect to the parameters that affect the computational budget and the number of dimensions of the search space. The parameters that affect the computational budget of the method are the number of



(a) 3 dimensions (b) 19 dimensions

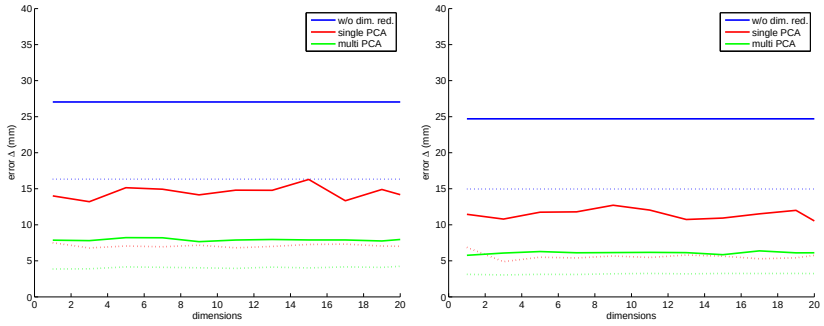
Fig. 3: Error Δ with respect to the number of PSO generations and particles: solid line for 16 particles, dotted line for 64 particles. PCA algorithms ran in (a) 3 PCA dimensions and (b) 19 PCA dimensions.

particles and generations of the PSO. More particles per generation search more densely the search space, while more generations improve the possibility that the particles converge and the global optimum is found. PCA dimensions can vary from 1 to 20. When no dimensionality reduction is used (baseline method), the articulation dimensions are 20. In both cases, the dimensions encoding position and orientation are 7 and will not be displayed in the following figures, even though they are subject to optimization.

In Fig. 3 the effect of the computational budget on the algorithms is shown. A first observation is that Δ decreases monotonically with the number of generations. Additionally, as the particles per generation increase, the resulting error decreases. However, multi-PCA seems to reach its performance peak at $20 - 25$ generations, which is sooner than the others. The results demonstrate that multi-PCA performs better than single-PCA, which in turn outperforms the baseline method. Furthermore, single-PCA with 16 PSO particles performs better that the baseline method with 64 PSO particles. Similarly, multi-PCA employing 16 particles has almost the same accuracy with single-PCA using 64 particles. Moreover, for an accuracy around $15mm$, the baseline method requires at least 30 PSO generations, while single-PCA requires only 16.

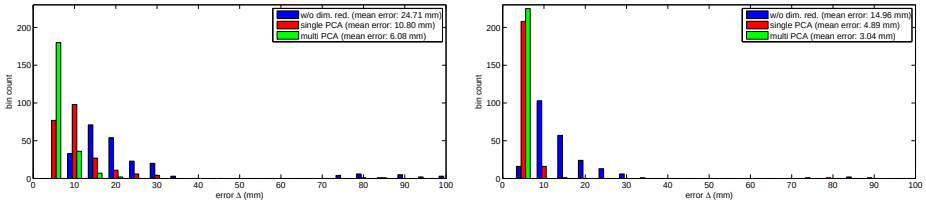
In Fig. 4 the horizontal axis refers to the number of PCA dimensions. The results show that the number of dimensions has a low impact on the single-PCA algorithm for the current dataset. It performs better in three dimensions than one, which was expected from the representation error analysis in Fig. 2. However, the best choice on the number of dimensions for a given accuracy cannot be safely estimated. The multi-PCA algorithm is less affected by the number of dimensions and approximates a straight horizontal line in the plots.

In order to examine the distribution of the error, the histogram of the estimation error Δ for all 3×75 estimations, using 35 PSO generations and 3 PCA dimensions, has been computed and is shown in Fig. 5. When using 16



(a) 20 generations (b) 35 generations

Fig. 4: The error Δ for the the 3 algorithms with respect to the number of PCA dimensions and the number of particles: 16 particles drawn as solid line, 64 particles drawn as dotted line. PSO generations are (a) 20 and (b) 35.



(a) 16 PSO particles (b) 64 PSO particles

Fig. 5: Histogram of the error Δ for all the 3×75 estimations per algorithm running for 35 PSO generations with (a) 16 PSO particles and (b) 64 PSO particles. Single-PCA and multi-PCA algorithms used 3 PCA dimensions.

particles, the baseline method has a small number of successful estimations on the $10mm$ error bin. Some estimations with error around $80mm$ correspond to wrong orientation estimation and usually a mirrored pose was returned. Single-PCA performs significantly better but with most of its estimations lying in the $10mm$ bin, it cannot be considered quite accurate. Multi-PCA performs exceptionally well, having 80% of its estimations in the $5mm$ bin, and has increased accuracy in more difficult poses.

When 64 particles are employed for each PSO generation, the baseline method's performance is improved but still 47% of the estimations exhibit an error greater than $12.5mm$. Single-PCA manages to almost reach the performance levels of the multi-PCA algorithm, while the latter could not make use of the extra computational budget since its performance was already very good. Figure 6 provides sample results obtained from the baseline and the two proposed methods for various PSO budgets.

Fig. 6: Sample hand pose estimations for no dimensionality reduction (rows 1, 2), single PCA (rows 3, 4) and multi PCA (rows 5, 6) rendered in blue. True poses are rendered in red color. For each method, the algorithms ran for 35 generations and for 16 (top) and 64 (bottom) particles. PCA algorithms used 3 dimensions.

4 Conclusions

In this work the usage of PCA in single frame hand pose estimation was assessed. The proposed single-PCA algorithm, given a dataset consisting of hand poses, calculates a space S_{7+M} of reduced dimensionality and implicitly learns some of the underlying hand configuration constraints. As indicated by the experimental evaluation, the results outperform the baseline algorithm. The estimated pose has an accuracy up to $5mm$ depending on the employed computational budget, and can be used to bootstrap automatically hand pose tracking. PCA, being a linear technique, cannot effectively describe complicated datasets with non linear data. To face this fact, we also proposed a multi-PCA algorithm which uses more than one training sets to learn multiple PCA subspaces. Multi-PCA performed better than single-PCA and required a quarter of the PSO particles to achieve slightly better accuracy. The obtained results demonstrate that model based methods can efficiently solve the single frame hand pose estimation problem in spaces of reduced dimensionality. This lifts one of their drawbacks, that is the need of manual initialization. Thus, their practical exploitation in the context of vision systems and applications is improved considerably.

Acknowledgments

This work was supported by the EU FP7-ICT-2011-9-601165 project WEARHAP.

References

1. R. Bowden, T. Heap, and C. Hart, "Virtual data gloves: Interacting with virtual environments through computer vision," *Proc. 3rd UK VR-Sig Conference, DeMontfort University, Leicester, UK*, 1996.
2. M. de La Gorce, N. Paragios, and D. J. Fleet, "Model-based hand tracking with texture, shading and self-occlusions," in *IEEE CVPR*. IEEE, 2008, pp. 1–8.
3. A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *CVIU*, vol. 108, no. 1-2, pp. 52–73, Oct. 2007.
4. M. Gorce, N. Paragios, and D. Fleet, "Model-Based Hand Tracking with Texture, Shading and Self-occlusions," *IEEE CVPR*, 2008.
5. H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool, "Tracking a hand manipulating an object," in *IEEE ICCV*. IEEE, 2009, pp. 1475–1482.
6. J. Ingram, K. Kording, I. Howard, and D. Wolpert, "The statistics of natural hand movements," *Experimental Brain Research*, vol. 288, pp. 223–236, 2008.
7. M. Kato, Y. wei Chen, and G. Xu, "Articulated hand tracking by pca-ica approach," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, April, pp. 329–334.
8. J. Kennedy and R. Eberhart, "Particle swarm optimization," in *IEEE ICNN*, vol. 4. IEEE, 1995, pp. 1942–1948.
9. J. Lin, Y. Wu, and T. S. Huang, "Modeling the constraints of human hand motion," in *Workshop on Human Motion, 2000*. IEEE, 2000, pp. 121–126.
10. I. Oikonomidis, N. Kyriazis, and A. Argyros, "Efficient model-based 3D tracking of hand articulations using Kinect," *BMVC*, 2011.
11. I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints," in *IEEE ICCV*. IEEE, 2011, pp. 2088–2095.
12. I. Oikonomidis, N. Kyriazis, and A. Argyros, "Markerless and efficient 26-dof hand pose recovery," *ACCV*, pp. 744–757, 2011.
13. P. Padeleris, X. Zabulis, and A. A. Argyros, "Head pose estimation on depth data based on Particle Swarm Optimization," *IEEE CVPRW*, pp. 42–49, Jun. 2012.
14. R. Poppe, "Vision-based human motion analysis: An overview," *CVIU*, vol. 108, no. 1, pp. 4–18, 2007.
15. J. Romero, H. Kjellstrom, and D. Kragic, "Monocular real-time 3d articulated hand pose estimation," in *Humanoid Robots, 2009. Humanoids 2009. 9th IEEE-RAS International Conference on*. IEEE, 2009, pp. 87–92.
16. R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff, "3d hand pose reconstruction using specialized mappings," in *IEEE ICCV*, vol. 1. IEEE, 2001, pp. 378–385.
17. M. Santello, M. Flanders, and J. Soechting, "Patterns of hand motion during grasping and the influence of sensory guidance," *Journal of Neuroscience*, vol. 22, pp. 1426–1435, 2002.
18. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," *IEEE CVPR*, pp. 1297–1304, Jun. 2011.
19. B. Stenger, P. R. Mendonça, and R. Cipolla, "Model-based 3d tracking of an articulated hand," in *IEEE CVPR*, vol. 2. IEEE, 2001, pp. II–310.
20. Y. Wu, T. S. Huang, and T. S. Huang, "View-independent Recognition of Hand Postures," in *IEEE CVPR*, 2000, pp. 88–94.