Inferring Hand Pose: A Comparative Study of Visual Shape Features

Akshaya Thippur Carl Henrik Ek Hedvig Kjellström

CVAP/CAS, KTH, Stockholm, Sweden

akshaya,chek,hedvig@kth.se

***Abstract*— Hand pose estimation from video is essential for a number of applications such as automatic sign language recognition and robot learning from demonstration. However, hand pose estimation is made difficult by the high degree of articulation of the hand; a realistic hand model is described with at least 35 dimensions, which means that it can assume a wide variety of poses, and there is a very high degree of self occlusion for most poses. Furthermore, different parts of the hand display very similar visual appearance; it is difficult to tell fingers apart in video. These properties of hands put hard requirements on visual features used for hand pose estimation and tracking. In this paper, we evaluate three different state-of-the-art visual shape descriptors, which are commonly used for hand and human body pose estimation. We study the nature of the mappings from the hand pose space to the feature spaces spanned by the visual descriptors, in terms of the smoothness, discriminability, and generativity of the pose-feature mappings, as well as their robustness to noise in terms of these properties. Based on this, we give recommendations on in which types of applications each visual shape descriptor is suitable.**

I. INTRODUCTION

Humans convey and relay a significant amount of information through non-verbal communication. An important part of this communication is carried out by moving our hands. To that end, being able to automatically estimate the movement and pose of hands [1] is very important for a range of applications such as robot learning from demonstration [2] and automatic sign language recognition [3].

However, the human hand is capable of a large range of poses and complicated movements. This places significant demands on the sensory system employed; it needs to be non-intrusive to not restrict the movement, but still needs to be able to reflect the intricate details of the hand. These requirements are met by a passive vision sensor, which does not require any equipment to be worn by the human, and is able to capture data at a frequency such that even the smallest changes in pose would be reflected. In this paper, we concentrate on visual hand pose estimation.

The mapping of a hand pose to its appearance in an image is very complex. Firstly, a hand is a very high-dimensional structure, which makes the range of poses that the hand can assume very large. Secondly, many parts of the hand, e.g., different fingers, have very similar appearance, which make them easily confusable. Furthermore, highly articulated structures like hands (and also full human bodies) display a great deal of self occlusion in the image. This means that many aspects of the hand pose are unobservable and have to be inferred from other aspects of the hand pose. Moreover, most information contained in an image are not relevant for pose estimation, but rather reflect orthogonal variations such as skin color and background patterns. This means that inference using such data is very challenging, as only a subset of the variations are correlated with the information we wish to infer; the pose, or functions thereof, such as the sign class in automatic sign language recognition. This is further formalized in Section II.

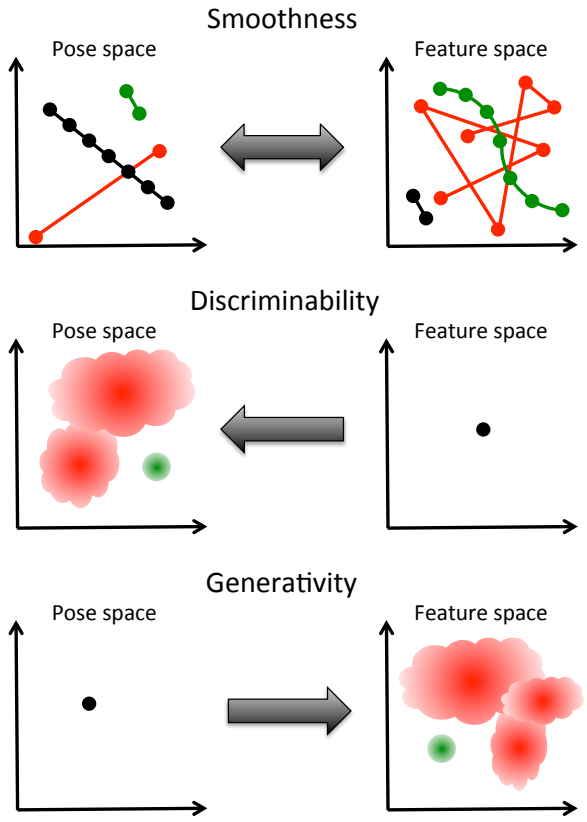


Fig. 1. Three desirable properties of features for hand pose estimation: **Smoothness:** A sequence of small motions in the pose space should lead to (●) a sequence of small motions in the feature space, not (●) a sequence of large motions. Likewise, the pose estimate should (●) be robust to small changes in feature value, not (●) change abruptly when the feature undergoes a small variation. **Discriminability:** A certain observed feature value should map to (●) a tight and unimodal distribution of poses, not to (●) a wide and multimodal distribution. **Generativity:** A certain pose should map to (●) a tight and unimodal distribution of generated feature values, not to (●) a wide and multimodal distribution.

To address this issue and simplify hand pose estimation from images, a number of shape features have been designed, which strive to extract the relevant information from

the image, and filter away the image variations that are not correlated with the hand pose. It would in theory be possible to learn a "perfect" hand feature (which reflected only the pose-related information) if one had access to enough training data. However, due to the extremely high dimensionality of image parameterizations, doing this in a principled manner would require an in-feasibly big data-set. To that end, most features are instead constructed in an ad-hoc manner, motivated by experimental results and intuitions rather than learned from training data.

Fig. 1 illustrates the desirable properties of successful visual shape features. First of all, for the estimation to be robust to noise, the mapping between a pose and its observed feature value should be smooth. This means that a small change in pose should not cause a great variation in feature value, and vice-versa, that a small random variation in the observed feature value (due to image noise) should not cause a great change in estimated pose. Secondly, if the feature is to be used for regression, it is important that it is possible to estimate the underlying pose from an observed feature value. This property can be denoted *discriminability*. Third, if the feature is to be used in a generative method such as an HMM or a particle filter, it is important that the feature generation process can be modeled in an accurate manner, i.e., that a certain pose always give rise to the same feature value, with small random variations due to image noise. This property can be denoted *generativity*.

The literature in computer vision and pattern recognition describes many different image features. However, it is often not completely understood which feature to choose for a specific context. In this paper we will perform a thorough evaluation of three popular shape based descriptors in the context of human pose estimation. In Section III we describe the three popular image features, while Section IV details how the test data are generated, and Section V describes the experiments. Our aim is to provide an intuition not only on *when* and for *what* scenarios the descriptors work but more importantly *why* they produce the results they do. This study forms a basis both for designing new descriptors and for choosing the right descriptor for a given task.

We do not aim go give a comprehensive overview of the state of the art in hand tracking, but rather give a systematic comparison of a representative set of shape features. For a recent review of hand tracking efforts, see [1].

II. HAND POSE ESTIMATION

We now formalize the hand pose estimation task. Let hand pose (or any other state defining the hand pose, e.g., sign type in a sign language application) be denoted X, and let an image observation of this hand pose (an image, or a functional descriptor extracted from the image, see Section III) be denoted Y. The process whereby images are generated from hands in the world can be described by the functional mapping

$$Y = f(X) + \nu(X) \tag{1}$$

where f is a deterministic function and ν is a noise term (arising from, e.g., noise when capturing the image, or from spurious changes in lighting conditions). This can be expressed in a probabilistic manner as if the observation Y is sampled from the likelihood of observations given the state X, $p(Y \mid X)$.[1] The function f and the statistics of the noise term ν, alternatively the likelihood density, can be learned from training data tuples $[X^i, Y^i]$.

The inference problem is to model the inverse of the image generating function – inferring the underlying state (hand pose) X from a certain observation (image descriptor) Y. Deterministically speaking, the inference problem boils down to modeling the inverse mapping f^{-1}. An approximation of this mapping can be learnt from data, if the true mapping f is one-to-one and the noise term ν is not too large. In probabilistic terms, the problem is to model the posterior density $p(X \mid Y)$, the probability density of the underlying state X given a certain observation Y.

For realistic applications it is not possible to learn the posterior directly, since the state space X is usually large, and the mapping f is usually many-to-many, making f^{-1} ill defined on the whole state space. There are two ways to go about in practice [4], the generative and discriminative approach. Both make use of the fact that whereas f^{-1} is ill-defined globally, it can be approximated locally around a certain prior estimate of X.

In the generative approach, used e.g. in [5], the posterior is modeled as a product of the likelihood and the prior over X using Bayes' theorem: $p(X \mid Y) \propto p(Y \mid X)p(X)$. The discriminative approach, used e.g. in [6], [7], models the likelihood $p(Y \mid X)$, not caring about how the states are distributed, and finds the X that maximizes the likelihood for a certain observed Y, using some optimization procedure.

Regardless of estimation or inference method, the image descriptor giving the measure Y should be such that the mapping f is as "nice" as possible – that is, contain as few singlarities or near-singularities as possible. In the optimal case, there is no observation noise ν, and the mapping function is unity: $Y = X$, i.e., the state (hand pose) is directly observable.

However, this is never the case in reality, as the hand pose space is very high-dimensional, and the hand pose X changes fast and in a non-linear manner over time, and since the image generating process (Section III) is highy complex. In this study, we evaluate three state-of-the-art visual hand shape descriptors with respect to

- A the smoothness of f – a small motion in the pose space X should correspond to a small motion in the feature space Y and vice versa, see Section V-A,
- B how functional f is – is there a clear unimodal maximum in the posterior distribution $p(X \mid Y)$ for a certain Y (discriminability), and is there a clear unimodal maximum in the likelihood $p(Y \mid X)$ for a certain X (generativity), see Section V-B,
- C how resistible the descriptor is to image noise,

[1] The notation is simplified, in that X indicates both the probabilistic state variable X and a certain value of this probabilistic variable, and Y denotes both the probabilistic observation variable Y and a certain observed value.

in terms of generativity and discriminability, see Section V-C,

and discuss them in relation to the needs of different applications of hand pose estimation.

Hand pose is almost always represented in terms of finger joint angles [1]. It should however be noted that this parameterization of hand pose displays certain peculiarities [8], [7]. For example, the Euclidean distances between two poses with change in the outermost joints might be larger than the Euclidean distance between two hands rotated 45° relative to each other – even though the latter two hands display vastly more different image appearances. Hauberg et al. [8] suggest an alternative state space with more coherence between pose difference and perceived image difference. However, the differences between these two pose spaces are not larger than that the findings in our study here are relevant to hand pose estimation in both spaces.

III. VISUAL HAND SHAPE DESCRIPTORS

The most straight-forward way to model the likelihood function $p(X \mid Y)$ is by using a articulated model of the hand, placing this hand model in the pose X and projecting it into the image. The appearance descriptor Y is some aspect of the image, e.g., the edges [6], [5] or the edge and ridge response [9]. In [10], the 2D finger positions in the image are estimated from the silhouette, which further guides the 3D hand model. The likelihood is measured as a function of the distance between the real image appearance Y and the image appearance postulated by the hand model in pose X. In the case of edges [6], the distance is in terms of the closest distance between the model and real edges, in the case of interest points [10], the distance is in terms of the closest distance between the model and the point detections, and in the case of filter responses [9], the distance is in terms of the actual filter responses at the model edges and ridges. If the measurement includes depth [11] or measurements from multiple cameras [12], the comparison can be made in terms of 3D hull instead of edges, which gives a more accurate likelihood estimate.

These model-based likelihood measures give very rich information about the hand appearance – hence their popularity for accurate pose estimation. However, they suffer from the short range of the measure in pose space: the estimate of the likelihood is only valid in a very narrow proximity of the true pose X. As an example, if the pose of the hand is correct, but the hand is shifted 2 cm to the left, the real edges of the index finger would match the model edges of the middle finger, the real edges of the middle finger would match the model edges of the ring finger, etc – the likelihood for this erroneous pose would display a local maximum. Hence, the likelihood densities in methods employing these kinds of descriptors are highly multi-modal. To deal with this, the pose estimation methods using these descriptors involve advanced optimization procedures [6], [11], or sequential tracking with careful manual initialization [5].

However, for applications when the full 3D hand pose is not estimated (e.g., automatic sign language recognition, where the state X is a sign class or a vector of descriptors of scene class), or there is no accurate initial estimate of the hand/human state X, or when computing power limitations prevent the use of advanced optimization methods, a more global descriptor (with a more smooth likelihood distribution) of hand pose is needed. Such applications include pedestrian detection [13], automatic sign language recognition [3], and detection-based hand and body tracking [14], [15], [16], [7], [17], [18].

An alternative approach to using a well-behaved feature with a smooth likelihood function is to use a simpler shape feature with a less smooth likelihood function, and compensate for this using a more elaborate functionality for exploring the state space, e.g., using clustering followed by a decision forest methodology for local classification/regression [19]. The trade-off is here in terms of computational speed: while the introduction of more well-behaved shape feature functions will improve any pose estimation method, the choice of feature might be limited to very simple features due to real-time requirements and computational power limitations. On the other hand, the computational cost of regression is lower with a more well-behaved feature. This trade-off has to be decided upon in each application.

In this paper we evaluate three state-of-the-art hand shape descriptors in the literature: *Hu moments* [20], *shape context* [21], and *histograms of oriented gradients* (HOG) [22].

A. Hu Moments

The 7 Hu moments [20] constitute a rotation, mirroring and scale invariant representation of the shape of a 2D silhouette. The moments are different combinations of scale-normalized centralized moments of the shape. The feature $Y_t^{\mathbf{hu}} = [h_{1,t}, \ldots, h_{7,t}]$, the 7 Hu moments extracted from a hand contour $\mathcal{H}_t$ segmented from the image frame I_t.

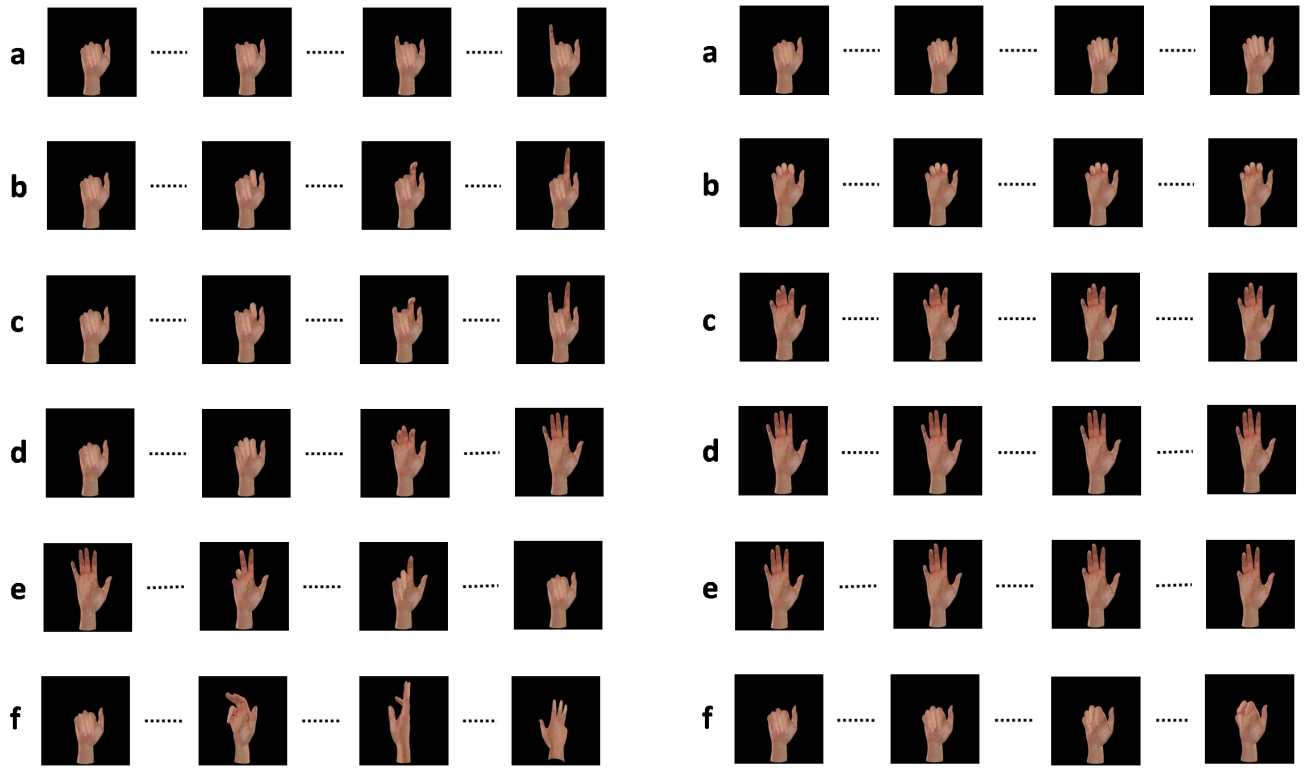
This image descriptor was commonly used for hand shape representation before 2005 [3], [1], [17]. After this, it has gradually been replaced by other shape descriptors – in Section V we show why this is the case.

This representation does not take any interior parts of the hand appearance into account, only the outer hand shape. One can thus assume that it is susceptible to errors in the segmentation of the hand from the background. Furthermore, since the representation includes higher-order derivatives, it is sensitive to small variations in shape.

B. Shape Context

Like Hu moments, shape context [21] is computed from the silhouette of a shape. However, while Hu moments are a global representation of the shape, a shape context is a local representation of a certain point on the silhouette. The shape context descriptor is a polar histogram around a certain edge point, where each bin counts the number of other edge points in this area of the neighborhood.

Shape context was originally designed for shape matching, i.e., matching certain points on one shape to the corresponding points on another shape To obtain a global shape descriptor – which will characterize the overall shape rather



(a) 6 large scale motion sequences

(b) 6 small scale motion sequences

Fig. 2. The 12 synthetic motion sequences. Each sequence corresponds to 100 equally long steps along a straight line in the hand pose space. For each sequence, frames 1, 34, 67 and 100 are shown. (a) Large scale motion. (b) Small scale motion.

than a certain point in relation to the rest of the shape – we use a bag of words approach: We learn a vocabulary V of size 128 of shape contexts from a representative set of contexts C extracted from the training data (silhouettes $\mathcal{H}$ from all hand frames I). To represent a hand shape (a silhouette $\mathcal{H}_t$ from the hand frame I_t), 100 points are sampled uniformly around the outer contour of the hand. Shape contexts $S_{i,t}$ (8×5 log-polar histograms) are computed for each point, and classified as words in the learnt vocabulary using a nearest-neighbor classifier $W = w(S)$ in the shape context space. The overall shape is then represented with a histogram H_t over all shape context words $W_{i,t}$. The feature $Y_t^{\mathbf{sc}} = [H_{1,t}, \ldots, H_{128,t}]$.

This representation has been used for regression-based hand and body pose estimation [14], [16] and has been proven highly discriminative of articulated pose.

Compared to Hu moments, the shape context descriptor is more robust to small variations in shape thanks to the binning and vector quantization steps. Therefore, one can expect it to have better smoothness and discriminability properties, but not necessarily to be more generative than Hu moments.

C. Histograms of Oriented Gradients

Unlike Hu moments and shape context, the histogram of oriented gradients (HOG) [22] descriptor operates on the entire hand image window I_t. Similarly to the SIFT feature [23], he image window I_t is partitioned into cells – here 5×7 cells – and for each cell, a histogram of gradient orientations – here with 12 bins – is computed. The feature $Y_t^{\mathbf{hog}} = [G_{1,t}, \ldots, G_{420,t}]$ where G_t is the concatenation of the gradient orientation histograms from all 35 cells.

The size of the cells and the granularity of the histograms affect the generalization capabilities of the feature. A more detailed discussion on how different parameters of the HOG affect human detection can be found in [22].

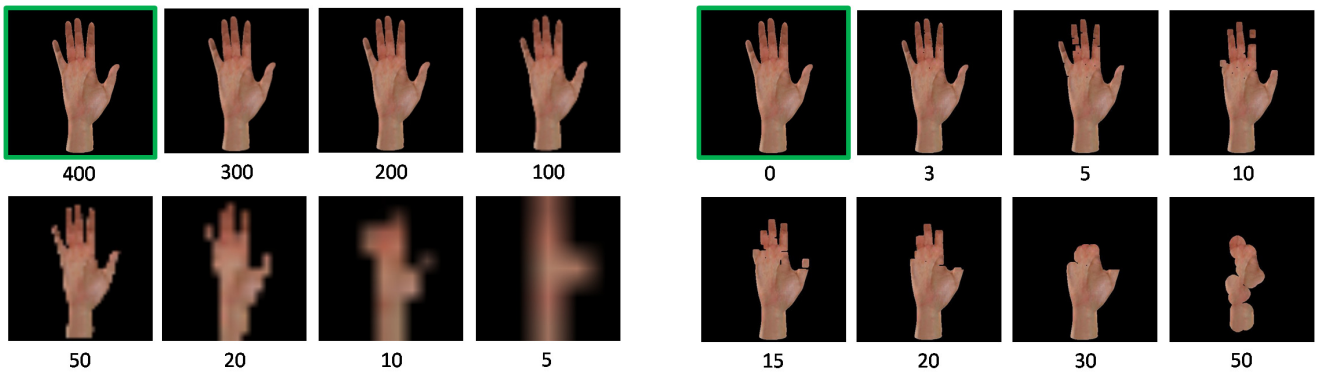
Due to its spatially coarser binning, the HOG descriptor is expected to be less discriminative than shape context. However, the robustness of the feature will probably lead to a very smooth mapping with good generativity – the feature value can be predicted in an accurate and deterministic manner from a hand pose.

HOG descriptors are mostly used for detection [22], [24] but also for regression-based hand and body pose estimation [25], [26], [7], [18], [27].

IV. DATA SET

To evaluate the three features described above, we need a set of sequences of hand views over time, together with their ground thruth pose. Ideally, we would evaluate on real image sequences, but given the difficulties of capturing ground truth pose without compromising the appearance of the hand, we generate a synthetic dataset using the *LibHand Library* [28].

The hand model used in LibHand has 63 degrees of freedom, defined by 54 joint angles and 9 camera orientation parameters, and can generate realistic 400×400 pixel image views of the hand in a given pose.



(a) Resolution noise

(b) Segmentation noise

Fig. 3. Simulated image noise. (a) Simulating the image quality loss that one gets when the hand is captured at low resolution: 400×400 pixels (original size, 0% noise), 300×300 pixels (scaled up to 400), 200×200 pixels (scaled up to 400), 100×100 pixels (scaled up to 400), 50×50 pixels (scaled up to 400), 20×20 pixels (scaled up to 400), 10×10 pixels (scaled up to 400), 5×5 pixels (scaled up to 400). (b) Simulating the image quality loss caused by errors in segmentation of the hand from the background: 0%, 3%, 5%, 10%, 15%, 20%, 30%, 50%.

A. *Sequences*

Fig. reffig:sequences shows the 12 sequences that are generated for these experiments. The 6 sequences in Fig. reffig:sequences(a) correspond to a large change in at least one of the pose parameters, while the 6 sequences in Fig. reffig:sequences(b) correspond to very small pose change. Each sequence, of length 100, correspond to 100 equally long steps along a trajectory in the hand pose space; in the case of **a**, **b**, **c**, **d**, a straight line. The camera orientation is kept fixed in all sequences.

For large scale motion, sequence

- **a** corresponds to straigtening the little finger,
- **b** corresponds to straightening the index finger,
- **c** corresponds to straightening the two fingers,
- **d** corresponds to opening the hand,
- **e** corresponds to closing the hand,
- **f** corresponds to opening and turning the hand.

The small scale sequences are parts of the large scale ones, with denser sampling.

B. *Simulating Image Noise*

To evaluate the effect of image noise, the synthesized hand views are perturbed in a controlled manner.

A common problem in many applications where the whole upper body is in view, such as automatic sign language recognition and human-robot interaction, is that the resolution is limited. To simulate this situation, 7 lower-resolution datasets are created where all hand views are down sampled, and then up-sampled to 400×400 again. Example images from the 7 datasets with different levels of resolution noise are shown in Fig. 3(a).

Another source of image noise is the segmentation of the hand from the background. This is typically done using skin color models, depth boundaries, motion, or a combination thereof. Segmentation results often suffer from errors along the boundary, or whole parts of the foreground missing. To simulate errors in the hand segmentation, 7 datasets with segmentation errors are created by diluting the original segmentation masks with random changes of pixels as fg/bg along the boundaries, followed by erosion and dilation. Example images from the 7 datasets with different levels of segmentation noise are shown in Fig. 3(b).

V. EVALUATION

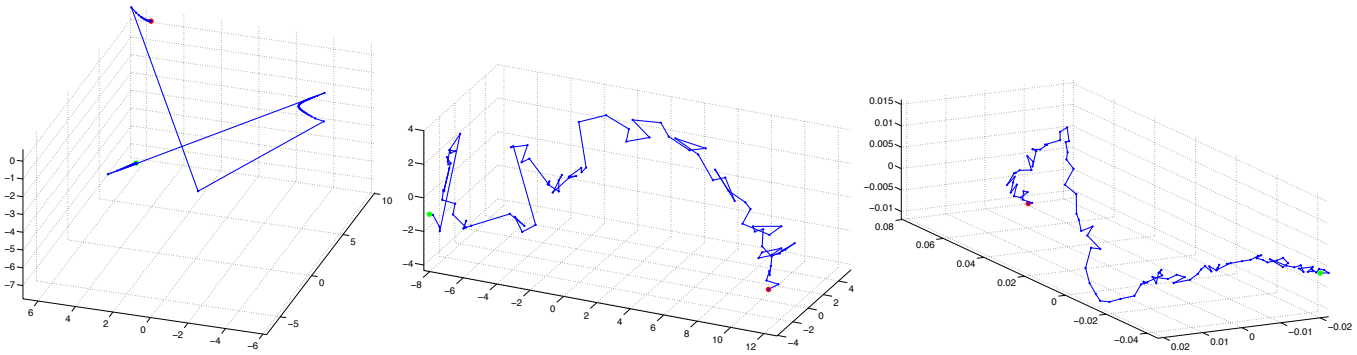
Using the hand sequences described above, the three features are now evaluated in terms of the smoothness, discriminability and generativity of the pose-feature mapping, as well as its robustness to image noise. The extraction of all features are implemented in Matlab.

A. *Evaluation of Smoothness*

For the ideal pose-feature mapping $Y = X$ as discussed in Section II, a sequence of poses X_t along a straight line in the pose space would render a straight line of features Y_t in the feature space. To evaluate how close to ideal a feature is, we can map a straight line in the pose space and study its corresponding line in the feature space. This gives an idea of the *smoothness* of the mapping f.

For each of the three features, the trajectory Y_t in the feature space corresponding to Sequence **d** from Fig. 2(a) is generated. The trajectories, projected down to their respective 3 largest modes of variation, are shown in Fig. 4.

The Hu moments trajectory $Y_t^{\mathbf{hu}}$ (Fig. 4(a)) has local segments where the features change very slightly between time steps, followed by sudden very large jumps. Thus, the Hu moments descriptor is sometimes very brittle to small changes in pose; the mapping is not very smooth. The shape context trajectory $Y_t^{\mathbf{sc}}$ (Fig. 4(b)) is much more smooth, i.e., there are fewer and less large jumps. The HOG trajectory $Y_t^{\mathbf{hog}}$ (Fig. 4(c)) is even smoother. There are three almost 90 degree turns. They most probably correspond to the frames in which the fingers (who are straightened, see Fig. 2(a)) show up in new cells in the HOG window. This means that the orientation histograms in the newly occupied cells suddenly start to change after being constant, causing a sudden change of feature trajectory direction.



(a) Hu moments (b) Shape context (c) HOG

Fig. 4. Trajectory in the 3 largest eigendirections of variation in the feature space, corresponding to a linear motion in pose space (Sequence **d** from Fig. 2(a)). (a) A Hu moments trajectory $Y_t^{\mathbf{hu}}$. (b) A Shape context trajectory $Y_t^{\mathbf{sc}}$. (c) A HOG trajectory $Y_t^{\mathbf{hog}}$.

From this we can conclude that HOG is a very robust descriptor, whereas Hu moments are brittle.

B. Evaluation of Discriminability and Generativity

A way to study to what degree the pose-feature mapping is one-to-one is to look at the coherence between distances in pose space $\delta X = \|X_1 - X_2\|$ vs distances in feature space $\delta Y = \|Y_1 - Y_2\|$. This can be visualized by taking the pairwise distances in pose and feature space of all pairs of points in each sequence in the dataset. A perfectly coherent, smooth, one-to-one mapping with no noise, $Y = f(X)$, would generate a histogram where only one bin per row and column are populated. In reality, this is not the case, since there are ambiguities and noise in the mapping.

Figure 5 shows the histograms for all three features. For each feature, the large scale motion histogram was generated from pairwise distances from the large scale sequences, while the small scale motion histogram was generated from the small scale sequences. The small scale histograms (b), (d), (f) can be regarded as a zoomed view of the upper left corner of the respective large scale histograms (a), (c), (e).

The *discriminability* of a feature can be measured using the standard deviation over δX given a certain δY; this says how coherently the inferred pose varies wrt the observed feature. According to this measure, shape context is the most discriminative feature, with HOG being nearly as good.

The *generativity* of a feature can be measured using the standard deviation over δY given a certain δX; this says how coherently the observed feature varies wrt underlying pose. According to this measure, HOG is by far the most generative feature.

C. Evaluating Robustness to Image Noise

Up to now, we have studied feature extraction from high resolution images with perfect background segmentation. To study the robustness of the three features to these two types of commonly occurring image noise, the analysis in Section V-B is performed on the 14 noise-polluted datasets from Section IV-B. The standard deviations over δX and δY for each feature are averaged. The mean standard deviations give a measure of the overall discriminability and generativity of different features at different noise levels.

Figure 6(a) shows the mean standard deviation as a function of resolution noise level. All features maintain their performance down to resolutions of 100×100 pixels. After that, the generativity of shape context goes down while Hu moments and HOG are robust to scalings down to as little as 20×20 pixels.

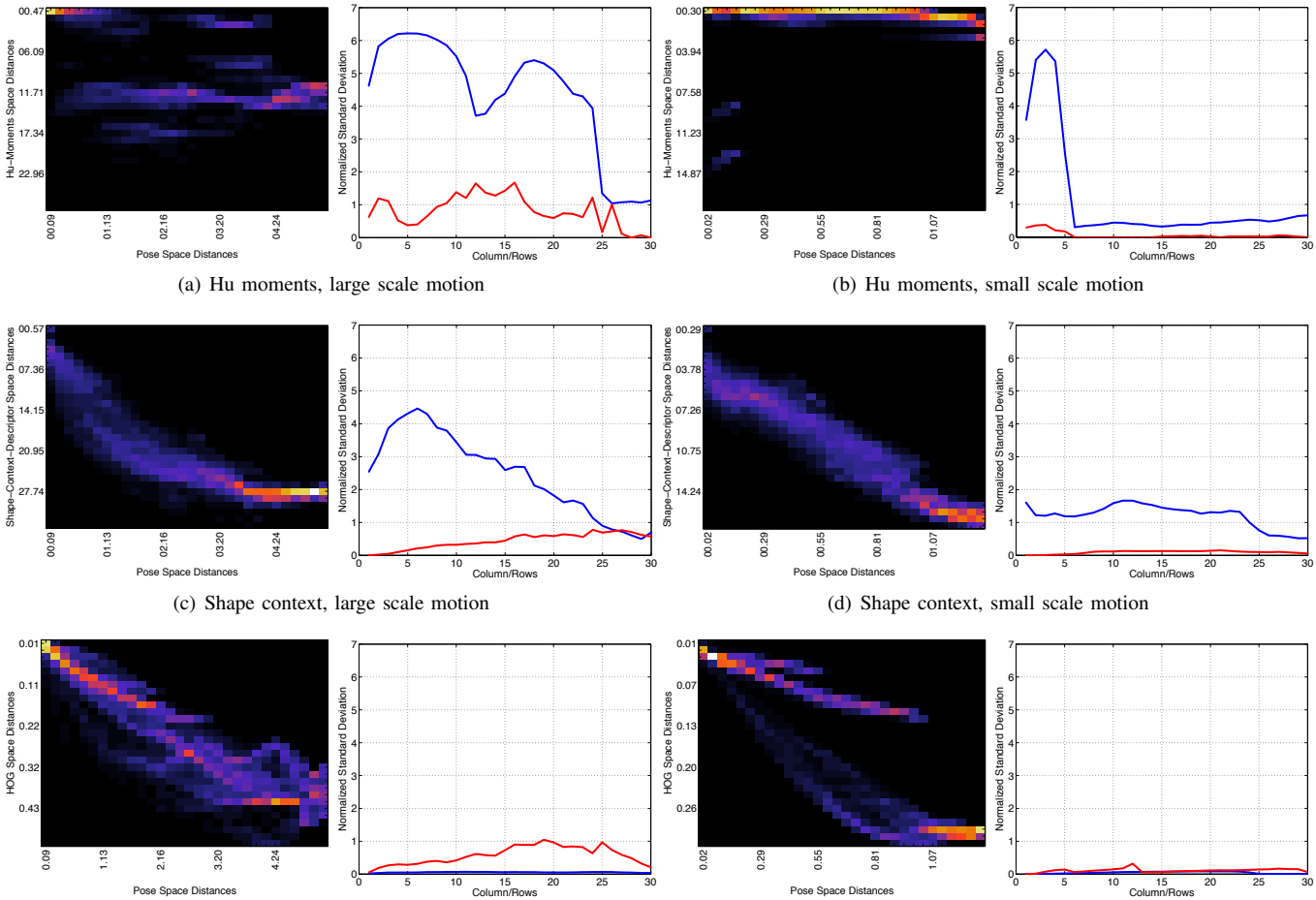
Figure 6(b) shows the mean standard deviation as a function of segmentation noise level. HOG is robust to up to 30% segmentation noise, while the others degrade already at 5%. The reason for this is that HOG uses not only the silhouette.

Figure 7 show how the HOG, large scale motion histogram deteriorates at the most severe noise levels. With low image resolution, the distances take discrete values, as many values as there are pixels on the hand. With noisy segmentation, the distances become uncorrelated with the underlying pose - this is natural since the pose is not visible even to a human looking at the image.

VI. CONCLUSIONS

We evaluate the three popular shape features Hu moments, shape context and HOG in terms of their efficiency for hand pose estimation. Based on our experiments, we can give the following recommendations:

- It is always better to use shape context or HOG instead of Hu moments, since these features reflect the pose in a more robust way.
- Shape context is very descriptive of hand pose, more so than HOG, which makes it the best choice for regression based methods. However, if there are large segmentation errors, HOG should be used instead, as it is more stable to this kind of image noise.
- HOG is the most generative feature, meaning that the HOG feature value can be predicted with high accuracy from a certain pose. This makes it suitable for generative estimation methods such as particle filters or HMM.
- HOG and Hu moments are resistible to low image resolution, which makes them preferable over shape



(a) Hu moments, large scale motion

(b) Hu moments, small scale motion

(c) Shape context, large scale motion

(d) Shape context, small scale motion

(e) HOG, large scale motion

(f) HOG, small scale motion

Fig. 5. Histograms of Euclidean distances in pose space $\delta X = \|X_1 - X_2\|$ vs Euclidean distances in feature space $\delta Y = \|Y_1 - Y_2\|$. Next to each histogram, the generativity, i.e., standard deviation over δY given a certain δX (•), and the discriminability, i.e., standard deviation over pose δX given a certain δY (•), are plotted. (a) Hu moments $\delta Y^{\mathbf{hu}}$, large scale motion. (b) Hu moments $\delta Y^{\mathbf{hu}}$, small scale motion. (c) Shape context $\delta Y^{\mathbf{sc}}$, large scale motion. (d) Shape context $\delta Y^{\mathbf{sc}}$, small scale motion. (e) HOG $\delta Y^{\mathbf{hog}}$, large scale motion. (f) HOG $\delta Y^{\mathbf{hog}}$, small scale motion.

context when the image resolution is low, e.g., for automatic sign language recognition.

Acknowledgments. This work was supported by the EU project TOMSY (ICT-FP7-270436).

REFERENCES

[1] A. Erol, B. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly. Vision-based hand pose estimation: A review. *CVIU*, 108(1–2):52–73, 2007.

[2] A. Billard, S. Calinon, R. Dillman, and S. Schaal. Robot programming by demonstration. In B. Siciliano and O. Khatib, editors, *Handbook of Robotics*, chapter 59. Springer, 2008.

[3] H. Cooper, B. Holt, and R. Bowden. Sign language recognition. In T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, editors, *Guide to Visual Analysis of Humans: Looking at People*, chapter 27. Springer, 2011.

[4] J. Laserre, C. M. Bishop, and T. M. Minka. Principled hybrids of generative and discriminative models. In *CVPR*, 2006.

[5] E. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky. Visual hand tracking using non-parametric belief propagation. In *IEEE Workshop on Generative Model Based Vision*, 2004.

[6] 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 *ICCV*, 2011.

[7] J. Romero, H. Kjellström, and D. Kragic. Hands in action: Real-time 3D reconstruction of hands in interaction with objects. In *ICRA*, 2010.

[8] S. Hauberg, S. Sommer, and K. Steenstrup Pedersen. Natural metrics and least-committed priors for articulated tracking. *IVC*, 30(6–7):453–461, 2012.

[9] H. Sidenbladh and M. J. Black. Learning the statistics of people in images and video. *IJCV*, 54(1/2/3):183–209, 2003.

[10] N. Shimada, Y. Shirai, Y. Kuno, and J. Miura. Hand gesture estimation and model refinement using monocular camera-ambiguity limitation by inequality constraints. In *FG*, 1998.

[11] A. Weiss, D. Hirshberg, and M. J. Black. Home 3D body scans from noisy image and range data. In *ICCV*, 2011.

[12] E. Ueda, Y. Matsumoto, M. Imai, and T. Ogasawara. A hand-pose estimation for vision-based human interfaces. *IEEE Transactions on Industrial Electronics*, 50(4):676–684, 2003.

[13] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, 2005.

[14] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *PAMI*, 28(1):44–58, 2006.

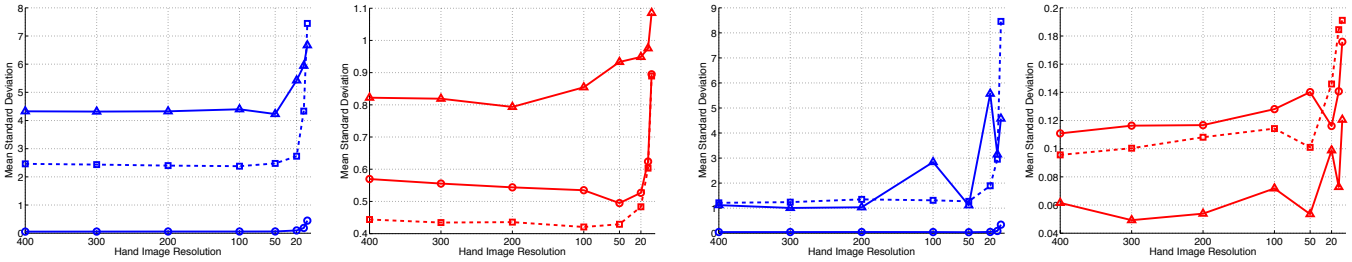
[15] V. Athitsos and S. Sclaroff. Estimating 3D hand pose from a cluttered image. In *CVPR*, 2003.

[16] T. E. de Campos and D. W. Murray. Regression-based hand pose estimation from multiple cameras. In *CVPR*, 2006.

[17] R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff. 3D hand pose reconstruction using specialized mappings. In *ICCV*, 2001.

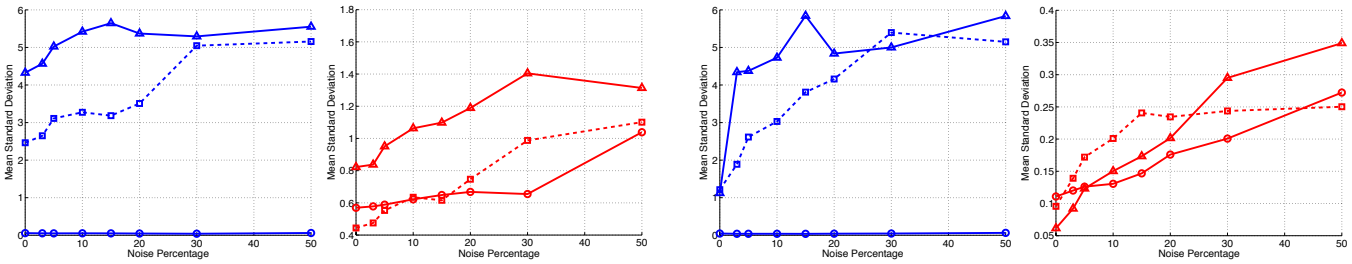
[18] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *ICCV*, 2003.

[19] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun. Hand pose estimation



(a) Resolution noise, large scale motion

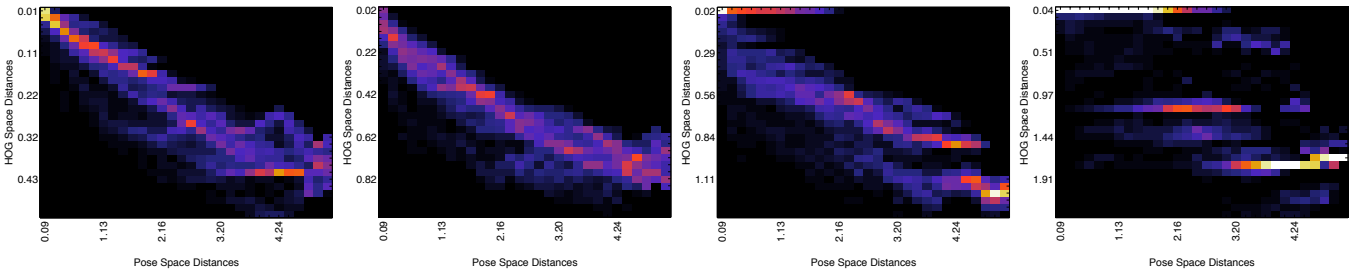
(b) Resolution noise, small scale motion



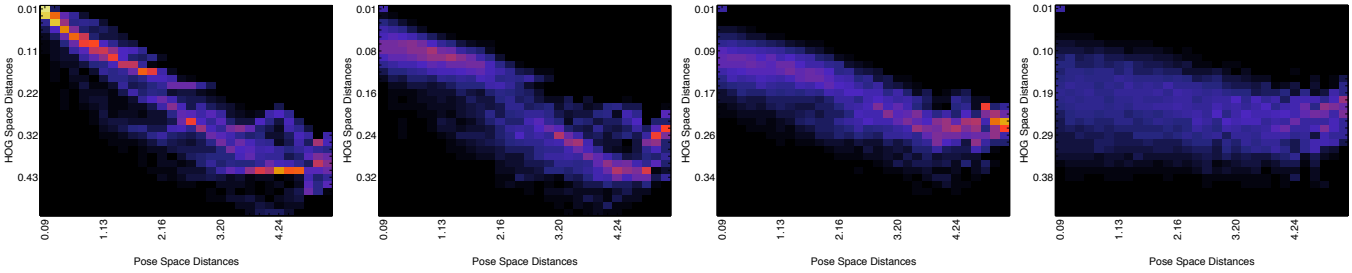
(c) Segmentation noise, large scale motion

(d) Segmentation noise, small scale motion

Fig. 6. Change in estimator accuracy with increasing image noise. To the left in each subfigure, the generativity, i.e., standard deviation over δY given a certain δX, averaged over δX (●) is plotted. To the right in each subfigure, the discriminability, i.e., standard deviation over pose δX given a certain δY, averaged over δY (●) is plotted. Solid curves with triangles = Hu moments, dashed curves with squares = shape context, solid curves with circles = HOG (a) Resolution noise, large scale motion. (b) Resolution noise, small scale motion. (c) Segmentation noise, large scale motion. (d) Segmentation noise, small scale motion.



(a) Resolution noise: 400×400 pixels (0%), 20×20 pixels, 10×10 pixels, 5×5 pixels



(b) Segmentation noise: 0%, 20%, 30%, 50%

Fig. 7. Change in feature vs pose distance histogram with increasing image noise. The figures show how the HOG, large scale motion histogram (Fig. 5(e)) is diffused with increasing noise. (a) Resolution noise: 400×400 pixels (original size, 0% noise), 20×20 pixels (scaled up to 400), 10×10 pixels (scaled up to 400), 5×5 pixels (scaled up to 400). (b) Segmentation noise: 0%, 20%, 30%, 50%.

and hand shape classification using multi-layered randomized decision forests. In *ECCV*, 2012.

[20] M. K. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187, 1962.

[21] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522, 2002.

[22] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[24] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.

[25] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

[26] V. Ferrari, M. Martin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

[27] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

[28] M. Saric. Libhand: A library for hand articulation, 2011. Version 0.9.