Experiments on Recognising the Handshape in Blobs Extracted from Sign Language Videos

Ville Viitaniemi and Matti Karppa and Jorma Laaksonen
Department of Information and Computer Science
Aalto University School of Science, Finland Email: firstname.lastname@aalto.fi

***Abstract*—Handshape has an important role in sign languages. It would be inconceivable to try to understand sign language without recognising the handshapes. Over the years, numerous different approaches have been proposed for extracting the hand configuration information. The existing approaches for handshape recognition have problems especially with the huge sizes of modern linguistic corpora. Computationally expensive methods become easily infeasible with such large amounts of data. In this paper we examine the straightforward and efficient approach of recognising handshapes by our existing image category detection methodology, involving state-of-the-art local image descriptors. In the experiments the approach produces promising results. On the image feature side, we find that surprisingly complex hierarchical descriptors of shape primitive statistics provide the best overall performance in handshape recognition. The accuracy of feature-wise detections can be improved by fusing together several features. Considering the temporal succession of the hand blobs markedly improves the accuracy over detecting the handshape in each video frame in isolation.**

I. Introduction

Handshapes – sometimes also called hand configurations [1] or hand poses – are explicit articulations that human beings form with their hands. Handshapes are among the most prominent features of sign languages, possibly only surpassed by the movement of the hands in passing linguistic information. It is inconceivable to try to understand sign language without recognising the handshapes. Over the years, numerous different approaches have been proposed for extracting the hand configuration information. The problem of recognising the 3D shape from a 2D projection is intrinsically difficult and all the proposed solutions suffer from various drawbacks. Besides the imperfect quality of analysis results, the computational cost of the proposed methods is another source of concern. The current trend in sign language research is to collect and analyse large corpora of at least dozens or hundreds of hours of signing. Computationally expensive methods easily become infeasible with such large amounts of data.

In this paper we examine the straightforward and efficient approach of recognising—or detecting— the handshapes in a real-world video material by the means of our existing image category detection system. It has earlier been successfully used by us for diverse other tasks in visual analysis. The system employs the standard processing stages of feature extraction and supervised learning. This appearance-based approach is taken here into an extreme in terms of simplicity of explicit hand modelling: hands are treated just as skin-coloured blobs of pixels. The implicit models of handshapes are provided by annotated training examples.

Our approach is related to some other appearance-based methods. However, our work has many distinctive properties. First of all, we use handshape examples from real video as models as opposed to synthetic images. Secondly, our models consist of handshape classes instead of individual prototype hand images. Thirdly, there exists a direct connection with our shape classes to the sign language phonology as our video material originates from a video dictionary of sign language. The handshape classes correspond to handshapes that the creators of the dictionary have judged to be phonologically distinctive. Fourthly, the features we extract from hands make the study interesting. The shape primitive statistic descriptors are quite advanced and complex as such, and may facilitate recognising handshapes in a degree that would not be possible using simpler features, thereby opening a door for new application possibilities. This study provides us the knowledge what level of performance can be realised with our current image analysis methods. We can then speculate whether this performance level is useful for applications in sign language analysis. We also gain insight into the usefulness of various feature extraction techniques in the handshape recognition context.

The rest of the paper is organised as follows. Section II surveys related literature. Section III describes the sign language video material we use along with the applied pre-processing. Section IV describes the used image feature extraction methods. Section V includes the description of our experiments along with the analysis of the results. Final conclusions are drawn in Section VI.

II. Related Work

A coarse division into two classes can be made for the existing handshape recognition methods: methods based on three-dimensional models and appearance-based methods. The 3D methods work by creating synthetic images by projecting hypothesised hand configurations onto the image plane and the hypothesis is updated after evaluating its correspondence with the input image. For example, in [2], an approach based on online optimisation of the parameters of a synthetic 3D model is presented. In [3], the problem is modelled as a detection problem and the configuration of the 3D model is selected via a Bayesian hierarchical detection scheme. Among other problems, both of these exemplary models suffer from high computational cost. In [4], a 3D model was used in a Monte Carlo importance sampling setting, and it was suggested that the dimensionality of the model could be reduced efficiently.

The appearance-based methods only model the 2D appearance of the hand, usually without taking the underlying skeletal structure into account. The methods tend to be simpler and thus

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.446

computationally less expensive, making them more viable for processing large amounts of data. However, the appearance-based hand analysis methods in the literature usually concentrate on detection, description and tracking hands on a lower level, the handshape is seldom explicitly modelled. At least not on the level of detail required by phonological analysis of sign language. Exceptions include [5] where the handshape estimation is treated as a database search of synthetic hand images. Although a 3D model is used, its use is limited to an off-line setting and matching is performed in an appearance-based fashion. In [6], a method of modelling the contour of the hand, based on a generalisation of Hidden Markov Models, was proposed, but the method is computationally very demanding. In [7], depth cues are obtained by estimating the direction of the shadow cast by the hand. Although the algorithm was reported to be very fast, the parametrisation is too crude for sign language context. A purely appearance-based approach is presented in [8], based on matching images with Dynamic Time Warping and Longest Common Subsequence measures. The approach is quite simplistic, but the basic problem setting resembles a smaller scale version of ours.

A recent survey into different methods of hand gesture recognition, covering a wide range of methodology, including hand tracking and other parts of a recognition system, is given in [9]. An older survey focusing on the hand pose estimation can be found in [10]. It would appear that most relevant work in recent literature has focused on 3D models. However, the current 3D methods are computationally too intensive for large scale data processing or require data collection procedures beyond the use of a standard 2-dimensional video camera.

III. Data

Our experiments were performed on a set of hand blobs extracted from a number of sign language videos of the S-pot benchmark material [11] that consists of the material from the Suvi video dictionary of Finnish Sign Language[1]. The video material for the dictionary was recorded during the 1990s with an analogue Betacam camera and converted into digital format afterwards. The technical quality poses a challenge to automatic analysis as the videos suffer quite severely from limited resolution, motion blur and coding artifacts. However, the challenge is a realistic one and reflects the real world: the dictionary can be regarded as a modern electronic sign language resourcewhose video quality is perfectly sufficient for a naturalistic viewing experience by human users.

A subset of 300 citation form videos was chosen from the material by considering those videos that were tagged with a specific handshape in the expert-prepared indexing of the dictionary. Furthermore, we limited ourselves to videos where the signer to wear a long-sleeved shirt. The chosen subset of videos was manually annotated by us. Our annotations specify the frames where the dominating hand of the signer shows the handshape of the index tag. Each video of the subset is indexed by a single handshape. This means that in each video, the occurrences of all the other non-indexing handshape classes are ignored.

[1]Suvi, the Online dictionary of Finland's sign languages, *http://suvi.viittomat.net*. Published in 2003, an extended version published with a new interface in 2013

Isolated skin-coloured hand blobs were extracted from the videos using an ELM-based detector. Because of technically modest video quality, the skin detection results had to be smoothed rather strongly, resulting in somewhat approximate and spread out blob contours. Other than that, the detection works well for the material. Dominant hand (i.e. right hand for right-handed signers) blobs were extracted only from frames where the hand was separate from the head and the non-dominant hand. In cases where the signer was left-handed, the hand blobs were mirrored so that the hand blobs in the extracted set all look like right hands. The automatically extracted blobs were screened by a human viewer. Erroneous blobs (e.g. hands occluding the head) were removed and the wrongly detected left/right-handedness corrected by mirroring. In the experiments we used both the exact skin-coloured area of the blob, and alternatively also the rectangular patch surrounding the blob. Our data (videos, blobs, annotations and extracted features) are available upon request.

Based on the visual properties of the isolated hand blobs alone, not all the blobs could be assigned the handshape class they are annotated with, not even by a human. This is because in our annotation procedure the annotator is able to see the whole video sequence and note down the beginning and ending times of a certain handshape. All the frames between these times get annotated with the handshape even though visually the handshape could not be identified in some of these frames. This often occurs due to motion blur, but sometimes also because of self-occlusion or otherwise difficult viewing angles. The annotations are thus somewhat noisy due to the annotation process not marking all the handshapes and assigning handshape labels to visually non-recognisable blobs. Still, the connection between the annotations and the actual handshapes is definitely strong enough to make automatically replicating the annotations a very feasible goal.

We devised an experimental setup where the task is to replicate the manually assigned handshape labels to blobs. Based on the numbers of extracted blobs associated with each handshape, the experiment was limited to the subset of 12 shapes with the largest number of examples. Altogether Suvi indexing uses 29 handshape classes. Figure 1 shows the subset.

Fig. 1. The 12 handshape classes of Suvi studied in the experiments.

For the purpose of the experiments, the videos were partitioned approximately evenly into training and test sets. This partitioning was constructed so that both the training and test sets contain roughly the same number of dominant hand blobs of each of the 12 handshapes. Table I shows details of the

TABLE I. STATISTICS OF THE DATA.

Handshape	Training videos	Training blobs	Testing videos	Testing blobs	Total videos	Total blobs
1001	24	340	24	362	50	702
1021	13	148	13	163	26	311
1100	3	44	4	54	7	98
1110	3	63	3	83	6	146
1120	7	143	6	85	13	228
1130	9	144	8	112	17	256
1201	19	234	19	229	38	463
1301	6	70	4	57	10	127
1311	6	85	5	55	11	140
1501	11	167	10	138	21	305
1511	8	107	7	90	15	197
1631	3	38	2	39	5	77
total	154	2462	146	2278	300	4740

statistics of the partitioning. The bottom row "Total" includes also blobs that are not annotated with any handshape label. There are 1239 unannotated blobs among the 4740.

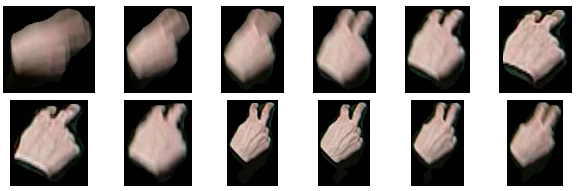

Fig. 2. Dominant hand blobs from a sequence annotated with label 1311.

IV. FEATURES

A large number of features and feature extraction variants was tested in the experiments with respect to their handshape distinguishing power. The features originate from our general purpose image/video analysis framework and have proven useful in diverse visual analysis tasks, including for example interactive multimedia retrieval and robot navigation [12].

Table II summarises the extracted features. Fourier descriptors and Zernike moments describe the shape of the extracted hand silhouettes. The remaining majority of the features describes the statistics of local shape primitives, mostly by means of histograms. The histograms are calculated either globally for the whole blob or the blob are first divided into sub-areas that are described each separately and the descriptions then joined. Different hierarchical blob area partitioning schemes result in a huge number of possible feature extraction variants. This number is increased by the independent design choice whether the features are extracted from the estimated skin area of the blobs or from the rectangular image patches surrounding the blobs (referred to as *patch* features from here on). The details of all the feature extraction variants we used are not described here due to space limitations, but some interesting issues are revisited in Section V-B where we analyse the experimental results.

Some of the features implement the bag-of-visual-words (BOV) feature extraction paradigm where codebook transform is applied to the image statistics. Within the paradigm, there are a few more parameters that can be altered: type of interest point selection (interest point detector or dense sampling), the size of the histograms (i.e. the number of histogram bins) and whether

TABLE II. EXTRACTED FEATURE TYPES.

Fourier descriptors of the blob contour
Zernike moments of the blob silhouette
Edge co-occurrence matrix
Edge histogram variants
Fourier transform of edges
Directional local brightness variation
spatial PCA of Census Transform histograms (sPACT) [13]
various Local Binary Pattern (LBP) histograms [14]
various Histogram of Oriented Gradients (HoG) features [15]
SIFT histograms with Harris-Laplace interest points (BOV)
ColorSIFT histograms employing dense sampling (BOV)

descriptors are assigned using soft or hard assignment. In our experiments we employ two BOV feature types: SIFT [16] and opponent colour ColorSIFT [17]. The sampling type is coupled with the descriptor type: the ColorSIFT features employ dense sampling whereas the SIFT features use Harris-Laplace interest point detection [18].

V. EXPERIMENTS

SVM classifiers were trained for all the different features and handshapes similarly as in the [12] (Section 3.3.1). The C-SVC soft margin variant was used, optimised with the LIBSVM library [19]. As kernels we used the χ^2 kernel for all the histogram-like features and RBF kernels for all the other features. A cross-validation type procedure was used for parameter search. , starting with a coarse line search, followed by a grid search for fine-tuning.

For each handshape, a classifier was trained using those blobs in the training half of the data for which the annotations specified the handshape to be definitely present or absent. The blobs with unknown annotation were excluded from training. The performance of the SVMs in recognising the handshapes of the test set blobs was evaluated in terms of average precision (AP). For each handshape, the evaluation was limited to those blobs of the testing half of the data for which the annotations definitely specified that handshape to be present or absent.

A. Detection of Individual Handshape Classes

The first 12 rows of Table III show the best recognition performance obtained for each handshape class. In the table, the columns under the heading "Best individual feature" address the best performance obtained by using any single one of the tested features. The column "AP" measures the absolute performance and "AP/a priori" shows the improvement over random (a priori) performance. By looking at these numbers, we see that the obtained recognition performance is clearly better than trivial, on average over four times as good. However, recognitions are far from perfect, as the AP values are clearly below one. It remains an open question, in which kinds of tasks is this performance level useful. In some applications handshape recognition that produces results of probabilistic nature can be useful already at the current level, e.g. in spotting specific signs within continuous signing [11]. However, if an application requires reliable crisp decisions of handshape, our current accuracy is probably not sufficient.

The row "MAP" of Table III shows the mean AP over all the 12 handshapes. On that row, the column "AP" shows the MAP of the single best feature, not the average of handshape-wise best performances. In addition, the row "MAP_4 shows

TABLE III. HANDSHAPE-WISE DETECTION RESULTS.

handshape	a priori	Best individual feature: best feature	AP	AP / a priori	Feature fusion: AP	Feature fusion: AP / a priori	Temporal smoothing: AP	Temporal smoothing: AP / a priori
1001	0.204	3×3 ColorSIFT, 512 bins	0.578	2.83	0.612	2.99	0.709	3.47
1021	0.093	Fourier descriptors, order 20	0.218	2.35	0.166	1.79	0.211	2.27
1100	0.031	SIFT, 4096 bins, patch	0.169	5.37	0.167	5.31	0.241	7.66
1110	0.048	3×3 HoG in spatial pyramid , patch	0.084	1.74	0.042	0.87	0.043	0.89
1120	0.049	sPACT, patch	0.097	1.97	0.075	1.52	0.101	2.05
1130	0.064	3×3 ColorSIFT, 512 bins, patch	0.245	3.82	0.297	4.63	0.314	4.90
1201	0.129	5×5 HoG	0.526	4.08	0.509	3.95	0.621	4.81
1301	0.033	3×3 HoG in spatial pyramid	0.157	4.79	0.151	4.61	0.174	5.31
1311	0.031	3×3 LBP	0.298	9.50	0.352	11.2	0.363	11.6
1501	0.079	Fourier descriptors, order 20	0.524	6.63	0.520	6.58	0.560	7.09
1511	0.052	1×1 HoG in spatial pyramid	0.258	4.98	0.243	4.69	0.304	5.87
1631	0.022	sPACT	0.461	20.4	0.282	12.5	0.428	18.9
MAP	0.070	5×5 HoG in spatial pyramid	0.235	3.37	0.284	4.08	0.339	4.86
MAP_4	0.126	5×5 HoG in spatial pyramid	0.429	3.40	0.452	3.58	0.525	4.16
MAP	0.070	Oracle feature selection	0.301	4.31				
MAP_4	0.126	Oracle feature selection	0.462	3.65				

the AP over the four classes that occur in more than 300 blobs in our data set (handshapes 1001, 1021, 1201 and 1501).

Figure 3 shows the best detection results for some handshapes. By inspecting similar visualisations for all the handshapes, we may state that for handshapes such as 1001, 1201 and 1501, the detections accuracy seems to be reasonable good. For example, within the 60 blobs detected as handshape 1001 with the most confidence, the "false positive" detections are actually blobs that just are not annotated with the handshape label in the ground truth despite showing the handshape. For some other handshapes, the results are modest at best. However, even though the detectors may not be able to capture the exact targeted Suvi handshape class, the best detections often seem reflect some other properties of the handshapes.

In a set of experiments we tried to improve the detection performance by combining the blob-wise detections based on a single feature in various ways. Firstly, we applied late fusion techniques to combine predictions made on basis of seven individually well-performing features: Fourier descriptors of orders up to 10 and 20, Zernike moments, global and 3×3 SIFT histograms and global and 3×3 ColorSIFT histograms. The results are shown in columns "Feature Fusion" in Table III. The columns "Temporal smoothing" refer to another series of experiments where the fused detections of the seven features were temporally spread onto preceding and subsequent blobs. In the experiments we tried both kernel smoothing and a maximum operation within a fixed-length window. Tabulated are the best results obtained using a Gaussian smoothing kernel with standard deviation $\sigma = 8$. The selection of the smoothing operator and its parameters are demonstrated more in detail in Figure 4. The results show that even this kind of elementary utilisation of the temporal dependencies of the hand blobs significantly improves the detection. The accuracy does not seem to be very sensitive to parameter selection. More advanced temporal modelling could probably improve the results even more.

B. Average Performance of Different Features

In the following we compare different features in the light of their average performance over the handshapes. Naturally, this overlooks the fact features compare differently for various handshape types, of which indications could be seen in the complete shapewise AP data (not shown here). However, we consider our sample size too small to look too much in the details.

Table IV displays the average performance of some of the features. Based on this table, we may make some observations regarding different feature extraction techniques. Firstly, we notice that features extracted from the exact skin area seem to usually perform somewhat better than features of the whole surrounding image patches. We did this comparison partly in the hope of being able to omit the sometimes inconvenient processing step of skin area determination and use just the surrounding patch, but this seems not to be a good choice. This is in contrast to the case in the task of more coarse blob type characterisation which we have previously investigated.

We also notice that sub-dividing the blob area into very small parts and describing each part in very much detail seems to lead to the best overall performance. For example, compare the performances of 1×1 LBP and 5×5 LBP. This is somewhat surprising, considering the small size of the blobs, the low image quality and the small training set. The best feature (5×5 HoG pyramid) employs a two-level spatial pyramid, each level consisting of HoGs evaluated in a 5×5 grid of 2×2 cells. This results in a 14580-dimensional feature vector. This is the most fine-grained HoG feature we evaluated, but finer partitionings might have worked even better. The nearly as good sPACT feature employs a three-level pyramid structure of census transform histograms (1302 dimensions). The exact way the partitioning is organised, hierarchically or otherwise, does not seem to matter as much as the sheer number of elementary parts. The line "edge co-occurrence matrix" is included in the table as an example of the much poorer performance of simpler shape statistic features.

The BOV features provide decent performance as well. Among them, ColorSIFT features perform better than the SIFT features in general, although shapewise there are exceptions. The features differ in two respects: the local descriptor type (SIFT or ColorSIFT) and type of point selection (Harris-Laplace interest point detector or dense sampling). The difference of the descriptors is the use of colour information in contrast to monochrome images. The colour probably does

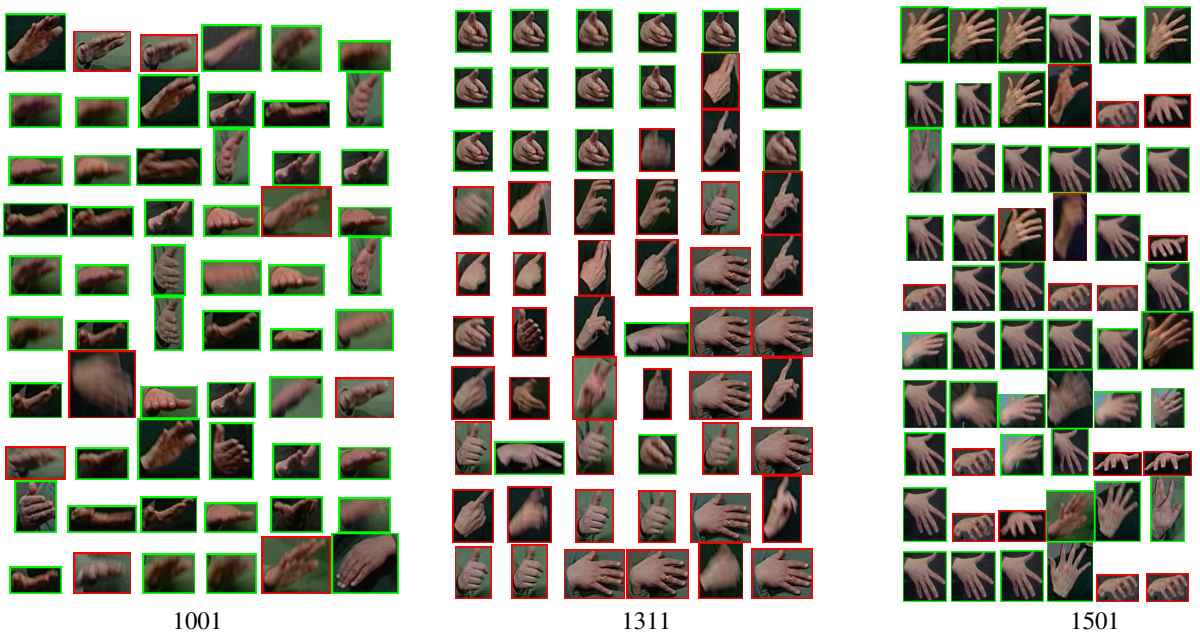

Fig. 3. The best 60 detections of handshapes 1001, 1311 and 1501. The green frames denote a correct detection according to the annotated ground truth, the red frames an incorrect one.

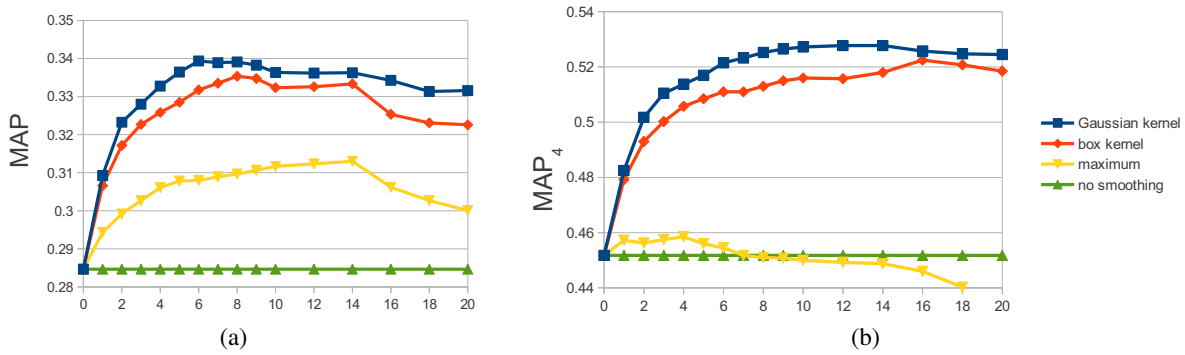

Fig. 4. The MAP (a) and MAP_4 (b) of different temporal smoothing operators and parameters. The temporal extent is the standard deviation of the Gaussian, the half-width of the box filter and the half-width of the maximum-taking window, in the case of Gaussian kernel, box kernel and maximum operator, respectively.

not differentiate the handshapes very much, which leaves us to conclude that dense sampling performs better overall than interest point detection in this handshape recognition task. However, based as well on the current experiments as on our prior experience from other applications, we may state that dense sampling and interest point detection are complementary: for some handshape and object classes interest point detection outperforms dense sampling. Also the BOV features are markedly improved by using sub-blob histograms with 3×3 partitioning of the hand blob area instead of single histograms for the whole blob.

The features characterising the shape of the hand silhouette, i.e. Fourier descriptors and Zernike moments, provide good performance. The Fourier features rival the best shape primitive statistic descriptors and outperform the BOV features. In light of more detailed handshape-wise results (omitted here), the features genuinely appear to be complementary in the sense that some of the handshapes can be recognised very well on basis of the blob contour shape descriptors whereas shape primitive statistics distinguish other shapes the best. The shape primitive statistics appear to provide steadier performance for all handshapes.

VI. Conclusions and Discussion

We have seen that with our current visual analysis and SVM detectors, we can achieve significantly better-than-random handshape recognition for hand blobs extracted from sign language videos. The quality of the results varies from handshape class to another. In light of examples, the system recognises some handshapes very well whereas the performance is not much better than random for some other shapes. It remains currently as an open question for what applications the reported detection accuracy is useful.

Whether or not the current machine learning setting is able to produce handshape detections that would be the most useful for practical applications, the experiments serve well in helping

TABLE IV. AVERAGE PERFORMACES OF DIFFERENT FEATURES.

		Skin area		Surrounding patch	
	Feature	MAP	MAP_4	MAP	MAP_4
Silhouette shape	Fourier descriptors, order 10	0.173	0.303		
	Fourier descriptors, order 20	0.210	0.390		
	Zernike moments	0.160	0.286		
Shape primitive statistics (non-BOV)	sPACT	0.234	0.395	0.207	0.370
	1×1 LBP	0.098	0.193	0.100	0.182
	3×3 LBP	0.198	0.342	0.162	0.284
	5×5 LBP	0.216	0.394	0.188	0.331
	pyramid LBP	0.190	0.349	0.159	0.286
	1×1 HoG	0.107	0.202	0.097	0.165
	1×1 HoG pyramid	0.190	0.309	0.127	0.235
	3×3 HoG	0.210	0.360	0.153	0.278
	3×3 HoG pyramid	0.230	0.378	0.167	0.328
	5×5 HoG	0.229	0.413	0.172	0.334
	5×5 HoG pyramid	0.235	0.429	0.174	0.326
	edge co-occurrence matrix	0.077	0.144	0.074	0.138
Shape primitive statistics (BOV)	ColorSIFT, 512 bins	0.180	0.310	0.143	0.265
	3×3 ColorSIFT, 512 bins	0.206	0.360	0.178	0.314
	SIFT, 4096 bins	0.159	0.283	0.151	0.272
	3×3 SIFT, 4096 bins	0.175	0.313	0.167	0.316

us understand how well we can expect to detect handshapes based on this type of visual analysis and how fine-grained distinctions we are going to be able to make. Furthermore, we now know more about the visual feature extraction methods that are useful for handshape recognition: surprisingly detailed statistical descriptions of the shape primitives within the blob area seem to work well. Limiting the descriptions to the actual skin area within the blobs improves handshape detection, in comparison to considering a rectangular image patch around the hand. In addition to shape primitive statistics, the shape of the hand silhouette is also useful. The accuracy of feature-wise detections can be improved by fusing together several features. Considering the temporal succession of the hand blobs significantly improves the accuracy over detecting the handshape in each video frame in isolation.

It may be that the taxonomy of handshapes used for indexing the Suvi dictionary is unnecessarily fine-grained for many practical applications of automatic handshape recognition, for example for automatic pairwise matching of sign sequences. It might be e.g. useful just to detect that the hand is compactly squeezed (the 1100-series) even if the exact configuration of the fingers could not be reliably determined, which would differentiate the handshapes 1100, 1110, 1120 and 1130. Likewise, detecting that fingers are spread out could be useful even if the handshapes 1501 and 1511 could not be told apart. Our future plans include a systematic experiment with the present hand blob collection to test which handshape groups could be separated most efficiently from others by automatic methods.

REFERENCES

[1] R. Johnson and S. Liddell, "Toward a phonetic representation of hand configuration: The fingers," *Sign Language Studies*, vol. 12, no. 1, pp. 5–45, 2011.

[2] M. de La Gorce, D. Fleet, and N. Paragios, "Model-based 3d hand pose estimation from monocular video," *IEEE PAMI*, vol. 33, no. 9, pp. 1793–1805, 2011.

[3] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla, "Model-based hand tracking using a hierarchical bayesian filter," *IEEE PAMI*, vol. 28, no. 9, pp. 1372–1384, 2006.

[4] Y. Wu, J. Lin, and T. Huang, "Capturing natural hand articulation," in *Proc. of IEEE ICCV 2001*, vol. 2, 2001, pp. 426–432.

[5] V. Athitsos and S. Sclaroff, "Estimating 3D hand pose from a cluttered image," in *Proc. of IEEE CVPR 2003*, vol. 2, 2003, pp. II–432–9.

[6] J. Wang, V. Athitsos, S. Sclaroff, and M. Betke, "Detecting objects of variable shape structure with hidden state shape models," *IEEE PAMI*, vol. 30, no. 3, pp. 477–492, 2008.

[7] J. Segen and S. Kumar, "Shadow gestures: 3d hand pose estimation using a single camera," in *Proc. of IEEE CVPR 1999*, vol. 1, 1999, pp. 479–485.

[8] A. Kuzmanic and V. Zanchi, "Hand shape classification using dtw and lcss as similarity measures for vision-based gesture recognition system," in *Proc. of EUROCON 2007*, 2007, pp. 264–269.

[9] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, pp. 1–54, 2012.

[10] 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, 2007, special Issue on Vision for Human-Computer Interaction.

[11] V. Viitaniemi, T. Jantunen, L. Savolainen, M. Karppa, and J. Laaksonen, "S-pot – a benchmark in spotting signs within continuous signing," in *Proc. of LREC 2014)*. Reykjavík, Iceland, May 2014.

[12] V. Viitaniemi, "Visual category detection: an experimental perspective," Ph.D. dissertation, Department of Information and Computer Science, Aalto University School of Science, May 2012, available online via http://lib.tkk.fi/Diss.

[13] J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE PAMI*, vol. 33, no. 8, pp. 1489–1501, 2011.

[14] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, January 1996.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR 2005*, vol. 1, 2005, pp. 886–893.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, November 2004.

[17] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluation of color descriptors for object and scene recognition," in *Proc. of IEEE CVPR 2008*, Anchorage, Alaska, USA, June 2008.

[18] K. Mikolajczyk and C. Schmid, "Scale and affine point invariant interest point detectors," *IJCV*, vol. 60, no. 1, pp. 68–86, 2004.

[19] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.