2D Appearance Based Techniques for Tracking the Signer Configuration in Sign Language Video Recordings*

Ville Viitaniemi, Matti Karppa, and Jorma Laaksonen

Department of Information and Computer Science,
Aalto University School of Science, Espoo, Finland
firstname.lastname@aalto.fi

Abstract. Current linguistic research on sign language is often based on analysing large corpora of video recordings. The videos must be annotated either manually or automatically. Automatic methods for estimating the signer body configuration—especially the hand positions and shapes—would thus be of great practical interest. Methods based on rigorous 3D and 2D modelling of the body parts have been presented. However, they face insurmountable problems of computational complexity due to the large sizes of modern linguistic corpora. In this paper we look at an alternative approach and investigate what can be achieved with the use of straightforward local 2D appearance based methods: template matching-based tracking of local image neighbourhoods and supervised skin blob category detection based on local appearance features. After describing these techniques, we construct a signer configuration estimation system using the described techniques among others, and demonstrate the system in the video material of Suvi dictionary of Finnish Sign Language.

1 Introduction

In automatic analysis of sign language videos it is necessary to detect and keep track of the configuration of the signer's body. On the coarsest level, the configuration modelling might correspond to roughly estimating the locations of the signer's hands. Estimates of handshapes, body poses, head poses and facial expressions add more details to the configuration modelling. However, estimating the configuration of the body, which is inherently three-dimensional, from a single 2D projection is not an easy task. The estimation can be done by fitting rendered complex 3D models to the observed image, as is done for example in [3] for a detailed model of the hand. Unfortunately these methods tend to be computationally far too expensive for amounts of video data that would need

* This work has been funded by the following grants of the Academy of Finland: 140245, Content-based video analysis and annotation of Finnish Sign Language (CoBaSiL); 251170, Finnish Centre of Excellence in Computational Inference Research (COIN).

to be analysed in the context of automatic analysis of sign language corpora of any practical significance. After all, there are 25 frames in every second of video footage.

Models based on *pictorial structures* [5] come closer to appearance-based modelling by operating directly in two dimensions. The model of [1] rather accurately manages to estimate the configuration of a 2D arm model with three joints—shoulders, elbows and wrists. However, estimating pictorial structure models is computationally quite heavy, too. The application of [1] to practical analysis is hindered also by another serious issue: the estimation needs extensive manually annotated training data, separately for each person to be modelled, and for each imaging condition, for example each clothing of the signer. [10] solves the issues of run-time computational complexity, but with the price of even more extreme training data requirements. Actually the method was demonstrated by training it with a huge video collection first processed with [1].

In this paper we go further in the 2D appearance based direction where the underlying models get simpler and simpler. The philosophy here is that simple methods should be preferred over more complicated ones if they are able to perform the same tasks. In this spirit we investigate whether techniques based on straightforward detection and tracking of local 2D appearance primitives would be sufficient for deducing the signer configuration in video material—at least in most cases. In such cases one could omit the more advanced and computationally complex modelling, which can be prohibitively cumbersome in practice.

This work studies the application of two kinds of local appearance based techniques. In the first technique (Section 2) we consider tracking of constellations of nearby small local image neighbourhoods. In sign language videos the objects of interest—primarily hands—are not rigid, but their shape and projection constantly changes. Furthermore, their appearance may vary also for technical reasons: in the video material we have cosidered, motion blur and coding artifacts are often quite severe. This quality is not unrealistically poor as the videos are from the Suvi online dictionary of Finnish Sign Language (FinSL) and offer perfectly naturalistic viewing experience for human viewers. The model of the object that is tracked thus consists of nearby image neighbourhoods that may gradually change their appearance from frame to frame and whose relative positions may also change. This technique was originally described in [16].

In the second technique (Section 3) we model the objects of interest as blobs of skin-coloured pixels. We categorise the skin blobs in the image into a small set of categories using standard off-the-shelf 2D visual descriptors combined with standard supervising learning algorithms: support vector machines (SVM) and extreme learning machines (ELM). In particular, we investigate whether these standard methods of 2D visual analysis can discriminate between right and left hand blobs, or tell which end of a blob depicting an arm is the end with the hand and the fingers. The left-right separation task has the practical motivation that in our video material, skin blob detection works rather well in identifying the skin areas that may be either hands or the face. In configuration estimation, the next task would then be to decide which skin blobs correspond

to which body part. Locating facial details is relatively easy, but hands are more problematic, especially deciding which hand is right and which one is left. In the video, the skin blobs can often be associated together by kinematic grounds when the movements are smooth, thus making the category estimation easier. A helpful clue is also provided by the areas where a skin blob moves: the right hand usually moves right of the left hand. However, these cues alone are not reliable enough to always correctly determine the configuration. Sometimes the movement is so fast that between subsequent frames the skin blobs have moved so much that associating the blobs in different frames is not straightforward. It may even be so that right hand appears nearly in the place where the left hand was previously and vice versa, throwing the kinematic analysis completely off the scent. Sometimes the image becomes temporarily unintelligible because of motion blur or the hands combine in a single blob with each other and/or the face, making it difficult to keep track of the position of each individual body part. Occasionally these factors combine. All in all, at times the tracking of configuration gets invalidated. Appearance-based blob categorisation can then come to help and re-establish the tracking.

After detailing the two techniques of main interest, we describe in Section 4 how we construct a system that performs signer body part identification and labelling in sign language videos. The system applies the two proposed techniques in several of its processing steps, as well as other image processing techniques. Section 5 demonstrates the system in labelling of videos of the Suvi dictionary. Section 6 discusses the findings and presents conclusions that can be drawn from this study.

2 Tracking of Local Image Neighbourhoods

Template matching based tracking of local image neighbourhoods—small rectangular patches in particular—is an important ingredient in our signer configuration tracking system. This elementary tracking method has been chosen because preliminary experiments with the video material at our disposal have indicated that some more advanced descriptors such as SIFT do not remain stable enough between the frames to make it possible to base the tracking on them. Partly this is because the appearance of the tracked objects keeps changing due to them being non-rigid, and also due to viewpoint changes. Image compression artifacts and motion blur are additional factors in our video material. In order to make the tracking more reliable, instead of tracking individual visual features we tie together a collection of multiple nearby points and track them collectively.

In the following we consider tracking a set of M points from a reference frame r to a target frame t. Let the coordinates of the tracked points be $\{\mathbf{r}_i\}_{i=1}^{M}$ in the reference frame. We impose a topology $\{N(i)\}_{i=1}^{M}$ upon the points. Here the set $N(i)$ specifies the indices of points that are neighbours of the ith point. There is no necessity of the topology to reflect any specific geometric notion of adjacency in the original image plane. In our formulation, the goal of the tracking is to find the coordinates $\{\mathbf{t}_i\}_{i=1}^{M}$ in the target frame so that the tracking cost function C

is minimised:

$$\min_{\{\mathbf{t}_i\}_{i=1}^{M}} C = \sum_{i=1}^{M} \left(A\big(I_r(\mathbf{r}_i), I_t(\mathbf{t}_i)\big) + \alpha \sum_{j \in N(i)} B\big(|\|\mathbf{t}_i - \mathbf{t}_j\| - \|\mathbf{r}_i - \mathbf{r}_j\||\big) \right). \quad (1)$$

Here $I_r(\mathbf{r}_i)$ and $I_t(\mathbf{t}_i)$ are the image neighbourhoods of the points $\mathbf{r}_i$ and $\mathbf{t}_i$ in the reference and target frames, respectively, and $A(\cdot,\cdot)$ is the template matching distance. α is a weight parameter of the method and $B(\cdot)$ is a scalar weighting function of distance differences. The cost function thus balances the sum of template matching costs of individual points with a measure how much the inter-point distances in the target frame differ from the corresponding distances in the reference frame. In this paper, the following iterative algorithm has been used for the approximate minimisation of the cost function:

1. Initialise tracking, i.e. select initial values for $\{\mathbf{t}_i\}_{i=1}^{M}$ e.g. on the basis of the estimated motion field.
2. Denote the set of indices of the target points requiring update with R. Initialise R by inserting all the indices $1, \ldots, M$ into it.
3. Repeat until R is empty or some external stopping criterion is met (e.g. number of iterations reaches a set maximum):
   (a) Randomly select an index j from R.
   (b) Set $\mathbf{t}_{\text{old}} = \mathbf{t}_j$ and remove j from R.
   (c) Search a new location for the point $\mathbf{t}_j$ that minimises C of Eq. (1).
   (d) If $\mathbf{t}_j \neq \mathbf{t}_{\text{old}}$, add indices in $N(j)$ into R.

3 Skin Blob Category Detection

For category detection, the object of interest is modelled as a blob, i.e. some contiguous sub-area of an image. Set of standard 2D visual features is extracted from the blob. Based on the features, the blob is assigned to one out of a pre-specified set of categories by supervised learning, i.e. using classifiers that have been trained previously with manually labelled data. In practice, the blobs we consider in our system result from skin colour detection. Here we consider assigning the blobs into two sets of categories that 1) distinguish between left and right hands, and 2) determine which end of an arm blob is the hand end. In [15] we have used similar methods for recognising pre-specified classes of handshapes.

The rest of this section describes an experiment that we performed in left-right hand separation in order to see which visual features are suitable as input to classifiers. Based on these results and those reported in [15], we have selected the sPACT statistical texture feature [19] for use in our system. In addition to classifier accuracy, also the computational lightness and ease of extracting the feature was a factor in the decision as we perform feature extraction and classification on-line for each video frame in our analysis system. For the same reason we also want to re-use the same features for all categorisation tasks, including handedness detection and handshape classification. sPACT provides decent performance in both tasks.

Table 1. The extracted features.

Shape	Fourier descriptors of the contour
	Zernike moments of the blob silhoutte
Texture statistics (BOV)	SIFT [7] histograms (Harris-Laplace interest points)
	Opponent colour ColorSIFT [13] histograms (dense sampling)
Texture statistics (non-BOV)	Edge co-occurrence matrix
	Edge histogram variants
	Fourier transform of edges
	Directional local brightness variation
	Spatial PCA of Census Transform histograms (sPACT) [19]
	Various Local Binary Pattern (LBP) histograms [9]
	Various Histogram of Oriented Gradients (HoG) features [2]

3.1 Right and Left Hand Discrimination Experiment

Our experiments were performed on a set of 7022 right hand blobs extracted from a number of sign language videos of the S-pot benchmark [14] that uses the material of the Suvi video dictionary [12] of Finnish Sign Language (FinSL). The blobs were mirrored to produce an artificial set of left hands. Half of the blobs was used for training and the other half for testing. In the experiment, a large set of visual features was extracted from the blobs (Table 1). A couple of features measure the shape of the extracted hand silhouettes (Fourier descriptors, Zernike moments). The remaining majority of features statistically describe the content of the blob area. When calculating statistical features, two independent design choices were explored: 1) extracting the feature from the actual skin blob area or its bounding box, and 2) calculating one feature vector for the whole blob or sub-dividing the blob area into parts and form the feature vector by concatenating the sub-part feature vectors.

In the experiment we systematically evaluated a large number of feature extraction parameter combinations. An SVM detector was trained for each of them, and the detector performance was then measured in the test set. For our on-line analysis system (Section 4), however, we use optimally pruned extreme learning machine (OP-ELM) classifiers [8] instead of SVMs because of the computational lightness of ELMs. Their classification accuracy almost equals that of SVMs. Table 2 highlights some of the SVM results. There we have included our best results for each feature type, as well as results that demonstrate some specific aspect of parameter selection. In the table, the first performance measure is the average precision (AP) in the task of ranking all the blobs in the test set according to their individual likelihood of showing a right hand. The pairwise error percentage measures the error in such a classification where the blob is compared with its mirror image and classified to be a right hand if the unmirrored blob has greater estimated right hand probability.

From the table we see that the handedness can be decided with a quite high accuracy based on the best of the blob appearance features. However, the accuracy drops markedly when instead of the complex texture features simpler elementary feature extraction methods are used (exemplified by the line “Edge

Table 2. Selection of left/right hand separation results.

Feature	skin/bbox	sub-divisions	AP	pairwise error-%	extraction time/ blob
random guess			0.5	50	
Fourier descriptors	skin		0.844	24.1	20 ms
Edge co-occurrence	skin		0.646	32.2	21 ms
matrix	bbox	5-part	0.852	21.1	22 ms
Edge	skin	4×4	0.907	14.8	18 ms
histogram	bbox	4×4	0.903	14.5	18 ms
	bbox	5-part $\otimes$ (4×4)	0.924	12.6	22 ms
sPACT	skin		0.966	9.6	11 ms
	bbox		0.977	7.1	11 ms
LBP	skin		0.873	18.4	8 ms
	bbox		0.890	18.2	8 ms
	bbox	5×5	0.980	7.3	9 ms
HoG	skin	1×1	0.884	17.1	13 ms
	bbox	1×1	0.883	17.8	13 ms
	bbox	2×2	0.950	10.0	13 ms
	bbox	$(2 \times 2) \otimes (2 \times 2)$	0.956	8.9	13 ms
ColorSIFT	bbox		0.975	7.3	380 ms
(512 bin codebook)	bbox	3×3 soft	0.983	6.2	390 ms

co-occurrence"). Overall the best performing feature is the most complex ColorSIFT feature we tested, the one with 3×3 soft spatial sub-division of the image area [17]. However, this feature implements the bag-of-visual-words (BOV) paradigm, making the feature extraction process computationally too heavy for on-line use in our system. We thus turn our attention to the best of other statistical texture features: LBP and sPACT. Subdividing the image area appears to be very beneficial in case of the LBP features (the sPACT feature already includes this internally). The 3×3 and 5×5 partitionings both appear to work well. Combining several different partitionings to a spatial pyramid does not bring further gains. The HoG features provide decent accuracy but are still clearly worse than LBP and sPACT. In case of LBP and sPACT, features extracted from the whole bounding box of the hand work better than limiting to the exact skin area of the hand. These results can be contrasted to the handshape recognition experiment [15] where HoG features extracted from the exact skin area worked best.

4 System for Signer Configuration Detection

We have devised a system for estimating the configuration of the signer in video recordings of sign language. More specifically, the system labels the skin pixels that correspond to the left and right hands of the signer and also identifies a single representative point for each hand. The videos have the constraint that they must portray a single signer filmed in a nearly frontal angle, without too

distracting background objects. In practice this is the video format used in the material of the Suvi dictionary. Our system utilises the techniques of Sections 2 and 3 among others and consists of the following processing steps:

1. Face detection with the Viola-Jones method [18]
2. Colour-based skin detection with an ELM detector
3. Overall skin blob progression tracking
4. Identification and elimination of unoccluded facial areas
5. Seed hand blob identification
6. Hand area tracking forward and backward in time from the seed blobs
7. Representative point selection for hand blobs
8. Normalisation of detected hand coordinates according to the shoulders

The system has been implemented in the SLMotion video analysis software framework [6]. Steps 1–4 refine the methods of [16] and are not described any closer here for space reasons. It may be said, however, that the steps 3 and 4 build heavily on point constellation tracking.

Hand seed selection in step 5 considers the hand blobs that are separate from head and the other hand and decides, which of the blobs corresponds to which hand. Seed selection divides into two cases by the sleeve length of the signer, which is detected automatically by considering the skin blob shape statistics. For short-sleeved signers, identification can be performed rather reliably on geometric grounds as the visible elbows constrain how far left the rightmost point in the right hand can be found, and similarly for the left hand. For long-sleeved signers, geometric cues are no longer reliable enough as only small areas of skin are visible. Hands often cross and the left hand appears right of the right hand. The decision of which hand is which is thus done by the methods of Section 3: sPACT features of the hand blobs are extracted and fed into pre-trained ELM classifiers. The temporal consistency of the handedness labelling is additionally improved by temporal smoothing.

In step 6 point constellations are selected from the identified seed hand blobs and tracked using the method of Section 2. The tracking is performed through the video both forward and backward in time until the next seed hand blob is encountered. The points that are tracked are selected among the output of a salient point detector [11], augmented with evenly spaced points in otherwise untracked areas. In a typical case approximately 100 constellations of 4 points each are tracked for each seed hand blob. After this step each detected skin pixel is labelled as either unoccluded face, right hand or left hand, based on the spatial distribution of the tracked point constellations.

Step 7 is performed differently for different sleeve lengths. For long-sleeved signers, the mass centre of the right hand pixels is used as the representative point for the right hand, and similarly for the left hand. For short-sleeved signers the hand blobs include also the visible arms and elbows, which makes the mass centre a bad choice. Instead, the hand end of the arm is identified with an ELM classifier based on the sPACT features (Section 3) and the reference point is chosen near the hand end. Before feature extraction, the principal axis of the arm is identified and rotated to be vertical.

In the final step 8 the coordinates of the labelled skin pixels and selected reference points of the hands are normalised within each video frame against translation and scaling by using a signer-centred coordinate system. To this end, the shoulder positions of the signer are estimated in each video frame. This involves colour-based torso mask estimation, shoulder template matching in the mask, and temporal filtering. The mean point between the two detected shoulders is chosen as the coordinate system origin and the inter-shoulder distance as the length unit.

5 Experimental Demonstration

We are routinely running versions of our system on the whole Suvi dictionary material (5539 videos) using a cluster of PC workstations. This shows that our analysis methods come to the level of computational complexity that they can be practically applied for relatively large linguistic corpora. Visual inspection of the results shows that in most cases, the hand detections are rather successful from the human observer's point of view. Figure 1 shows some detections. However, problems do sometimes occur in all the stages of the system, starting from the skin detections being erroneous. Generally, long-sleeved signers present more problems to our system than the short-sleeved ones.

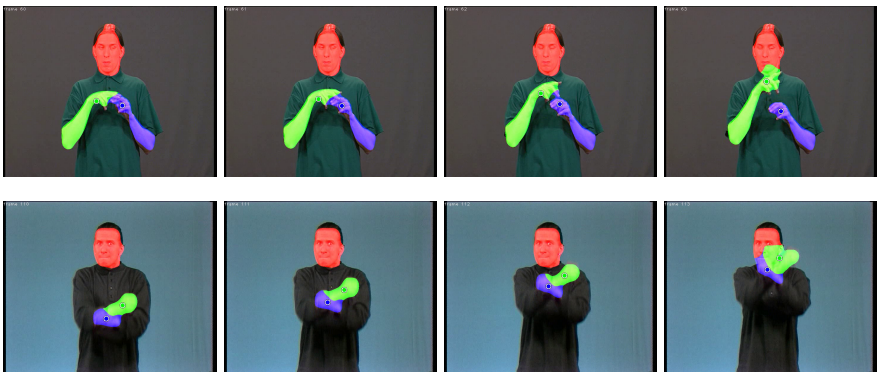

Fig. 1. Sample signer configuration detections in Suvi videos. The green mask denotes estimated right hand area, the blue mask the left hand area, and the red mask the unoccluded parts of face. The circles are the representative points selected for the hands.

Quantitative evaluation of the hand detection accuracy is challenging as we would need to measure the performance in a task that is known to be solvable well if the hand coordinates are known. Some benchmarking has been done [4] by the distance between automatic detections and manually estimated hand point locations. However, this task seems rather artificial in itself as it is difficult to

say, what would the accuracy level need to be in order it to be useful in solving practical tasks. In contrast, we have tried to solve the sign spotting benchmark task [14] using the automatically estimated hand coordinates. The problem here is that it is not yet known how well the task can be solved even with perfect hand location estimates. When we replaced the skin distribution histograms of our earlier DTW-based solution with the estimated hand coordinates, the accuracy remained on the same 48% level, which can be considered a good result since the two hand coordinate points are a much more compact representation of the signer configuration than the full spatial skin histograms.

6 Conclusions and Discussion

In this paper we have described two straightforward 2D appearance based techniques for tracking the signer configuration in video recordings of sign language: template matching-based tracking of local image neighbourhoods and skin blob category detection using appearance features and supervised learning. For the category detection, we have evaluated several types of visual features and found that rather complicated statistical texture features work the best, despite the typically small size of the skin blobs.

We described how to construct a system for signer configuration—here mainly the hand location—detection and tracking using the above mentioned techniques among others. The system has turned out to be practically feasible for investigating linguistic corpora of large size. The configuration detection accuracy is not perfect, but seems acceptable in the majority of cases. We plan to use the system in applications where it would not be important to get each and every detection absolutely right, but where information of more statistical nature would already be beneficial, such as information retrieval type of tasks. One can—for example—combine the hand detection with the handshape recognition techniques of [15] and look at the statistics of handshape distribution in signed material, which is a linguistically interesting question. On the other hand, the combination of hand locations and hand shapes might be used for addressing the sign spotting benchmark task of [14].

References

1. P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proceedings of the British Machine Vision Conference*, 2008.
2. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Procedings of Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.
3. M. de La Gorce, D. Fleet, and N. Paragios. Model-based 3D hand pose estimation from monocular video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(9):1793–1805, 2011.

4. P. Dreuw, J. Forster, and H. Ney. Tracking benchmark databases for video-based sign language recognition. In *Proceedings of the ECCV International Workshop on Sign, Gesture, and Activity (SGA)*, pages 286–297, Crete, Greece, September 2010.
5. P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61(1):55–79, Jan. 2005.
6. M. Karppa, V. Viitaniemi, M. Luzardo, J. Laaksonen, and T. Jantunen. SLMotion – an extensible sign language oriented video analysis tool. In *Proceedings of 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavík, Iceland, May 2014. European Language Resources Association.
7. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.
8. Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse. OP-ELM: Optimally-pruned extreme learning machine. *IEEE Transactions on Neural Networks*, 21(1):158–162, January 2010.
9. T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, January 1996.
10. T. Pfister, J. Charles, M. Everingham, and A. Zisserman. Automatic and efficient long term arm and hand tracking for continuous sign language TV broadcasts. In *British Machine Vision Conference*, 2012.
11. J. Shi and C. Tomasi. Good features to track. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '94)*, pages 593–600, June 1994.
12. Suvi, the on-line dictionary of Finnish Sign Language. http://suvi.viittomat.net, 2013. The online service was opened in 2003 and the user interface has been renewed in 2013.
13. 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.
14. V. Viitaniemi, T. Jantunen, L. Savolainen, M. Karppa, and J. Laaksonen. S-pot – a benchmark in spotting signs within continuous signing. In *Proceedings of 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavík, Iceland, May 2014. European Language Resources Association.
15. V. Viitaniemi, M. Karppa, and J. Laaksonen. Experiments on recognising the handshape in blobs extracted from sign language videos. In *Proceedings of 22th International Conference on Pattern Recognition (ICPR)*, Stockholm, Sweden, August 2014.
16. V. Viitaniemi, M. Karppa, J. Laaksonen, and T. Jantunen. Detecting hand-head occlusions in sign language video. In *Proceedings of the 18th Scandinavian Conference on Image Analysis*, volume 7944 of *LNCS*, Espoo, Finland, June 2013. Springer Verlag.
17. V. Viitaniemi and J. Laaksonen. Spatial extensions to bag of visual words. In *Proceedings of ACM International Conference on Image and Video Retrieval (CIVR 2009)*, Fira, Greece, July 2009.
18. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, pages I:511–518, 2001.
19. J. Wu and J. M. Rehg. CENTRIST: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1489–1501, 2011.