Long Term Arm and Hand Tracking for Continuous Sign Language TV Broadcasts

P. Buehler[1], M. Everingham[2], D. P. Huttenlocher[3] and A. Zisserman[1]

[1] Department of Engineering Science, University of Oxford, UK
[2] School of Computing, University of Leeds, UK
[3] Computer Science Department, Cornell University, USA

Abstract

The goal of this work is to detect hand and arm positions over continuous sign language video sequences of more than one hour in length.

We cast the problem as inference in a generative model of the image. Under this model, limb detection is expensive due to the very large number of possible configurations each part can assume. We make the following contributions to reduce this cost: (i) using efficient sampling from a pictorial structure proposal distribution to obtain reasonable configurations; (ii) identifying a large set of frames where correct configurations can be inferred, and using temporal tracking elsewhere.

Results are reported for signing footage with changing background, challenging image conditions, and different signers; and we show that the method is able to identify the true arm and hand locations. The results exceed the state-of-the-art for the length and stability of continuous limb tracking.

1 Introduction

The goal of this work is to find hand and arm positions in long sequences of continuous video. Our work is motivated by a long term goal of automatic sign language recognition, where knowledge of the hand position and shape is a pre-requisite. Our source material is the signing which typically accompanies TV broadcasts, such as BBC news footage or educational programs. This is very challenging material for a number of reasons, including self-occlusion of the signer, self-shadowing, motion blur due to the speed of motion, and in particular the changing background (since the signer is superimposed over the moving video). The difficulty of the video is illustrated in Figure 1.

Previous approaches have concentrated on obtaining the hands by using their skin colour [1, 3, 18] or by detectors based on AdaBoost hand classifiers [8, 13]. However, methods concentrating solely on the hands suffer when the hands overlap, or are in front of the head, and lose track due to the ambiguities that routinely arise. Ultimately, identifying the wrist position, hand angle, and assigning hands to be left or right with these approaches is not robust enough for reliable performance on long sequences.

The solution to these problems that we adopt here is to use the arms to disambiguate where the hands are. This is taking on a larger pose estimation problem (since now multiple limbs must be estimated) but in the end the constraint provided by the arms is worth the cost. To our knowledge no prior work in sign language adopts this approach.

Figure 1: (a-d) Image characteristics which render arm detection ambiguous: (a) Shading and self-occlusions significantly change the appearance of the left arm. (b) Similar foreground and background colours render the colour cue less informative. (c) Motion blur removes much of the edge and illumination gradients of the arms. (d) Proximity of the two hands makes the assignment into left and right hand ambiguous. (e) Upper body model used for pose detection.

There has been much previous work on 2D human pose estimation, mainly using pictorial structures [5, 6] based on tree structured graphical models. Their great advantage is the low complexity of inference [4] and hence they have been used in numerous applications [14, 15, 17]. While the run-time performance is very compelling, this approach has several limitations: (i) only pixels which are covered by the model contribute to the overall probability of a given limb configuration, i.e. any negative evidence (such as a skin coloured region in the background) is missed; (ii) over-counting of evidence, in that pixels can contribute more than once to the cost function and hence multiple parts can explain the same image area (as noted also by e.g. [16]). Finally, (iii) no modelling of occlusions, resulting in erroneous fits if a part is occluded.

We present a model which does not suffer from these problems. Every pixel in the image is generated using a limb model or the background model, taking proper account of self-occlusions. It is similar in form to that of [7, 9, 10, 11, 16] and described in detail in Section 2. Although this model satisfies our requirements, it is too expensive to fit exhaustively. Consequently, we propose inference methods which avoid such search, but achieve acceptable results: first, a sampling based method for single frames, where a modified pictorial structure is used as a proposal distribution (Section 3); and second, a method of identifying *distinctive frames* which can successfully be linked by simply tracking configurations (Section 4).

We evaluate our model and inference procedures on continuous signing footage taken from BBC broadcasts with changing backgrounds, using ground truth annotations. Results are reported in Section 5.

2 Generative Hand and Arm Model

This section describes our generative model which explains every pixel in the image and hence also takes the background as well as occlusions into account.

Complete cost function: Formally, given a rectangular sub-image $\mathbf{I}$ that contains the upper body of the person and background, we want to find the arm configuration $\mathbf{L} = (b, l_1, l_2, ..., l_n)$ which best explains the image, where $\{l_i\}$ specifies the parts (limbs) and b is a binary variable indicating the depth ordering of the two arms. In our application we deal with $n = 6$ parts: the left and right upper arms, the forearms and the hands. The position of the head, torso and shoulders are obtained in advance by a pre-process, and the appearance (e.g. colour) and shape of the parts are learned from manual annotation of a small number of training images (detailed in Section 2.1). The background is continuously varying, and largely unknown.

Every part $l_i = (s_i, \alpha_i)$ is specified by the two parameters scale (foreshortening) s_i and rotation α_i, and by the part to which it is connected . The connections are in the form of a kinematic chain for the left and right arm respectively (see Figure 1(e)). While the upper and lower arm can undergo foreshortening, the parameter s for the two hands is fixed.

We define the probability of a given configuration $\mathbf{L}$ conditioned on the image $\mathbf{I}$ to be

$$P(\mathbf{L}|\mathbf{I}) \propto P(\mathbf{L}) \prod_{i=1}^{N} p(x_i|\lambda_i) \prod_{j \in \{LF,RF\}} p(y_j|l_j) \quad (1)$$

where N is the number of pixels in the input image, x_i is the colour of pixel i, and y_j is the HOG descriptor computed for limb j (see Section 2.1).

The cost function incorporates two appearance terms which model the agreement between the image $\mathbf{I}$ and configuration $\mathbf{L}$. The first, $p(x_i|\lambda_i)$, models the likelihood of the observed pixel colours. Given the configuration $\mathbf{L}$, every pixel of the image is assigned a label $\lambda_i = \Lambda(\mathbf{L}, i)$ which selects which part of the model is to explain that pixel (background, torso, arm, etc.). The "labelling" function $\Lambda(\mathbf{L}, i)$ is defined algorithmically essentially by rendering the model (Figure 1(e)) in back-to-front depth order (the "painter's algorithm") such that occlusions are handled correctly. For a given pixel, the colour likelihood is defined according to the corresponding label (see Section 2.1). Note that the pixel-wise appearance term is defined over *all* pixels of the image, including background pixels not lying under any part of the model. The second appearance term, $p(y_j|l_j)$, models the likelihood of HOG descriptors extracted for the left and right forearms $\{LF, RF\}$. Implementation details for both likelihood terms are given in Section 2.1.

The third term, $P(\mathbf{L})$, models the prior probability of configuration $\mathbf{L}$. This places plausible limits on the joint angles of the hands relative to the lower arms.

Complexity of inference: There are 11 degrees of freedom: 5 for each arm and 1 for the depth ordering. The state spaces of the arm parts are discretised into 12 scales and 36 rotations. The hand angle is restricted to be within 50 degrees relative to the lower arm and discretised into 11 rotations. Hence, the total number of possible arm configurations is $2 \times ((12 \times 36)^2 \times 11)^2 \approx 10^{13}$. Brute force optimisation over such a large parameter space is not feasible – the method described in Section 3 addresses this problem.

2.1 Implementation details

This section discusses how the likelihoods are computed for a given configuration $\mathbf{L}$ (which in turn defines the pixel labelling). The extent of user input necessary to learn the model is also described.

Colour cue: The colour distributions for the foreground parts are modelled with mixtures of Gaussians using manually labelled data. Since the signer appears only in a corner of the employed TV footage, the background colour distribution is learned from the remaining image area. An RGB histogram representation is used, and updated every frame to account for the changing background. Given the learned distributions, every image pixel x_i is assigned a likelihood $p(x_i|\lambda_i)$ for each label λ_i.

Histogram of Gradients cue: In previous work the appearance of the different parts is often described using edge information at their boundaries. In our case these boundaries are not very reliable cues, due to motion blur, self-occlusions, shading, and strong folds in the clothing, as well as due to the sleeves having the same colour as the torso (see Figure 1). We exploit both boundary and internal features to determine the position and

configuration of a part using Histogram of Oriented Gradients (HOG) [2, 19] templates. The agreement between the template of a given part with configuration l_i, and the HOG descriptor of the image is evaluated using a distance function.

A HOG template is learned from manually labelled data for each part, scale and orientation. The individual templates are calculated as the mean over multiple training examples with similar scale and angle. If no or only few such examples exist, then additional examples are generated from the closest training instances by appropriate rotation and scaling.

The distance function for the HOG appearance term, $p(y_j|l_j)$, is computed by evaluating a truncated L2 norm and normalising it to be in the range $[0,1]$. The image area covered by a template could be used to scale the importance of the HOG cue, although we found that this degraded the results. In our experiments we only calculate the HOG appearance term for the left and right forearm since these are often not occluded, and provide the strongest constraint on the wrist position.

Hand labelling: Since the hands can assume many different shapes, using just a rectangular model would be an inaccurate representation. Instead, Λ is defined to label a pixel as hand only if it is contained in the rectangle corresponding to the hand part *and* if the pixel is skin-like (that is, if the colour is more likely to be hand than any other part).

Head and torso segmentation: In a similar manner to [12], the outline of the head and the torso are detected first before identifying the arm configuration. Multiple examples of the head and torso segments are used as templates (binary masks) and fitted to an image using a simple two part pictorial structure. Each part is specified by 4 degrees of freedom: translation in x and y, rotation and scale. The posterior distribution for this model is similar to (Eqn. 2) but here the appearance term uses only part-specific colour distributions, while the binary term, $P(\mathbf{L})$, acts as a spring-like force which enforces that the head is close to the position of the neck. The MAP estimate could then be used as a segmentation of the head and the body. However, this would restrict the shape each part can assume to the shapes given in the training examples. Instead, we (i) select all templates close to the MAP estimate, (ii) assign each template a weight given through the appearance term and (iii) calculate the final segmentation by applying a threshold to the weighted sum over all templates. The shoulder position is similarly estimated by a linear combination of the shoulder positions in each template.

Learning the model: Manually labelled data is required to learn part-specific colour distributions, to build the head and torso model, and to create HOG templates. The colour distributions were learned from only 5 manually segmented frames. For head and torso segmentation 20 examples were segmented and used as templates. HOG templates were learned from 39 images where the true arm configuration was manually specified. We achieved the best results using a HOG descriptor with cell size of 8×8 pixels, block size of 2×2 cells, and 6 orientation bins.

3 Computationally Efficient Model Fitting

As will be demonstrated in Section 5, the optimum of the complete cost function defined in the previous section correlates very well with the true arm configuration. However, the vast number of possible limb configurations makes exhaustive search for a global optimum of the complete cost function infeasible. In Section 3.1 we propose an effective approximation where the arms are fitted sequentially. Section 3.2 shows how this approximation can be combined with a fast approach based on sampling.

3.1 Iterative arm fitting

The configuration of the arms is estimated in two iterations, by fitting each arm in turn while holding the other fixed. First, the best location of the left arm is found, then the best location of the right arm while keeping the left arm fixed. The process is then repeated with the ordering reversed, that is the right arm is fitted first. The minimum cost configuration is then chosen as the final result.

Performing two passes in this way, with the arms fitted in each order, is essential to avoid ambiguities in the hand assignment. For example fixing the left arm can in some cases claim the right hand, and leave only unlikely positions for the right arm. When reversing the ordering, the right arm will claim its true location while also leaving the likely positions for the left arm. Our results indicate that this heuristic is a good approximation of the exhaustive search.

3.2 Sampling framework

Fitting the arms iteratively reduces the search complexity from $O(N^2)$ to $O(N)$ in the number of single arm configurations, but is still computationally expensive. We therefore adopt a *stochastic* search for each arm, using an efficient sampling method [4] to propose likely candidate configurations. Pictorial structures are well suited for this task: adopting a tree-like topology between connected parts, the posterior factors into unary and binary terms (Eqn. 2). However, this model has several shortcomings which were explained in the introduction, e.g. the susceptibility to over-count evidence. We show that by *combining* this sampling framework to hypothesise configurations, with our complete cost function to assess the quality of the sampled configurations, we obtain the robustness of our generative model with the computational efficiency of pictorial structures.

The posterior distribution from which samples are drawn is given [4] as

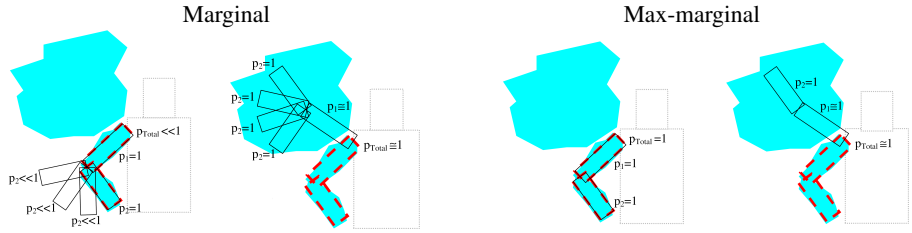
$$P(\mathbf{L}|\mathbf{I}) \propto P(\mathbf{L}) \prod_{i=1}^{n} p(z_i|l_i) \tag{2}$$

where $\mathbf{L} = (l_1, ..., l_n)$ defines the configuration of each part and z_i refers to the pixels covered by part i. $P(\mathbf{L})$ is defined as in Section 2. Samples can be drawn from (Eqn. 2) using the marginal distributions (see [4]). Alternatively, we argue in the next section that the use of max-marginals is better suited for this task (see Figure 2), where the summation operation of standard marginalisation is replaced by a maximisation operation. Formally, for a random variable $X = \{X_1, \ldots, X_n\}$, the marginal probability for a particular element X_i is defined as $P_m(X_i = k) = \int_{x:x_i=k} P(X = x)dx$. The max-marginal is obtained by replacing the integral with a max operation: $P_{mm}(X_i = k) = \max_{x:x_i=k} P(X = x)$. The significance is explained in the next section and Figure 2.

The appearance term, $p(z_i|l_i)$, is composed of the product of pixel likelihoods using colour distributions modelled by mixtures of Gaussians, and edge and illumination cues added through HOG descriptors (see Section 2.1 for details). In [4], the colour likelihood of a part is given as a function of the foreground and the surrounding background. This centre-surround approach is not suitable for our domain, where the arms are mostly in front of the torso, since the arm and torso are similar in colour.

3.3 Modifying the sampling

When using a sampling method to propose plausible arm locations, it is important that the true arm configuration is contained in the set of samples. In this respect the pictorial structure sampler is insufficient, for example given an image where a part is partially or



(a) true location (b) wrong location (c) true location (d) wrong location

Figure 2: This example illustrates that drawing samples of the upper arm from the max-marginal distribution can be superior to using the marginal. Figures (a) and (b) show two cases where the upper arm rectangle is either placed on the true location (red dotted line) or on a background area with arm-like colour (turquoise). The likelihood of the upper arm in isolation is equal in both positions. However, the marginal probability over the lower arm in (a) is low since only very few configurations exist which place the lower arm rectangle on the expected colour. This is in contrast to (b) where the marginal probability is high due to a large arm-like coloured area in the background. Hence, when sampling arm configurations using the marginal the upper arm will most frequently be sampled from the wrong image area (b). By contrast, the max-marginal for (c) and (d) is equal, since in both cases there is at least one lower arm position with high likelihood. Hence, by using the max-marginal, samples will be generated more often in the true arm location than using the marginal.

completely occluded, the associated probability for this part to be generated from its true location can be very low. We propose several extensions:

(i) Sampling from the max-marginal distribution: In contrast to common practice we sample from the max-marginal instead of the marginal distribution. Our reasoning is similar to that in [20], where the max-marginal is applied to problems where there are multiple possible valid solutions. The tree-model proposal distribution has precisely this property, due to issues such as the ambiguity caused by overlapping parts (see Figure 2).

(ii) Adding knowledge of the occlusion ordering: A part which is occluded has a very low probability to be proposed at its true position. However, in the signing scenario we know most of the occlusion ordering in advance: the arms are always in front of the torso and the hands are always in front of the arms. We can make use of this ordering by modifying the colour likelihood term: The likelihood for a given pixel is re-defined as the maximum likelihood for that pixel over the colour models corresponding to the part *and* all parts which could be occluding it according to the hypothesised configuration.

(iii) Sampling less often within the head and torso: If the sleeves and the torso share the same colour, many samples for the arms will be generated on the torso rather than on the true arm position. However, by knowing the outline of the torso (Section 2.1) we can "bias" the sampler to generate more samples outside the torso. This is achieved by decreasing the colour likelihood within the torso region (by sharpening). Even if both arms lie on the torso, then given that the background does not contain a high proportion of sleeve-like colours, most samples will still be generated on the arms.

(iv) Sharpening instead of smoothing the probability distribution: In [4] the authors recommend that samples be drawn from a smoothed probability distribution. In this work, in combination with the extensions listed above, we found it to be more beneficial to sharpen the distribution (see Section 5) instead (that is to the take the distribution to the power of τ with $\tau > 1$, in contrast to smoothing where $\tau < 1$).

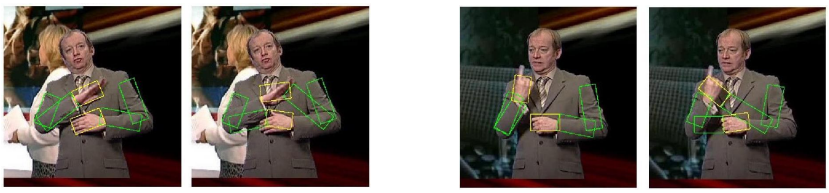


Figure 3: Identification of distinctive frames by exchanging the hand positions of the fitted model. (Left pair) Ambiguous: difference in cost is low. (Right pair) Distinctive: difference in cost is high.

4 Tracking using Distinctive Frames

While we have concentrated thus far on pose estimation in isolated frames, for video sequences it is valuable to exploit the temporal coherence between frames. Especially for ambiguous poses, which may introduce multiple modes in the complete cost function, this can significantly improve model fitting. We propose a method based on detection of unambiguous "distinctive" frames, and tracking between pairs of such frames. We show that distinctive frames can be detected not more than a few seconds apart, and hence losing track between such frames is not a problem. Our method owes some inspiration to [15], which detects a distinctive lateral walking pose to initialise a person-specific colour model. Our method differs in that the frequency of the detected frames allows the method to be used in a tracking framework, rather than solely for initialisation.

4.1 Identification of distinctive frames

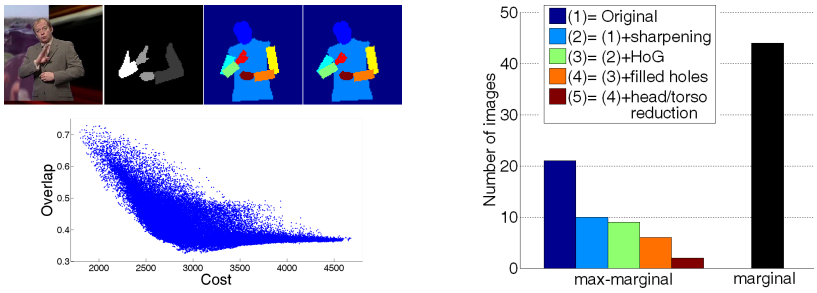
Our method of identifying distinctive frames builds on a set of rules obtained by analysing the complete cost function. We observe that most cases where the true pose is not identified are due to a confusion between the left and right hands or (less frequently) due to the background containing limb-like structures. Considering only frames where both hands are on the torso, pose errors are predominantly due to the problem of distinguishing the hands into left or right.

Using these observations, we define a simple heuristic to label a given frame as distinctive or ambiguous: (i) if two skin coloured areas are present on the body then (ii) find the arm configuration with minimum cost. If this configuration places the hands on the two skin-coloured areas, then (iii) perform a second inference step with swapped hand positions (see Figure 3). (iv) Mark the frame as distinctive if the difference in cost between the configurations from (ii) and (iii) is above a threshold.

4.2 Tracking between distinctive frames

Using the method of the previous section a large set of distinctive frames can be identified where the position of both arms is estimated with high accuracy. We now focus on finding the arm configuration for all remaining frames. This is implemented by tracking forwards and backwards in time between two neighbouring distinctive frames.

Temporal tracking is realised by adding the tracking term $P(\mathbf{L}|\mathbf{L}') = \prod_{k=1}^{n} p(l_k|l'_k)$ to the complete model in Eqn. 1 where $\mathbf{L}' = (b', l'_1, l'_2, ..., l'_n)$ refers to the part configurations in the preceding frame. The conditional probability $p(l_k|l'_k)$ is close to one if l_k and l'_k are similar and close or equal to zero for physically unrealistic cases, e.g. if the arm position changes dramatically within only one frame. We automatically learn a histogram representation of $p(l_k|l'_k)$ for each part, using a signing sequence where the background is static and the sleeves and the torso are of a different colour – for such a setting our approach gives very reliable results without the temporal term.



(a) (b)

Figure 4: (a, bottom) Complete cost vs. overlap with ground truth calculated by fixing the right arm to the location with maximum overlap, and evaluating all possible configurations for the left arm. (a, top) From left to right: Input image, manual ground truth, arm configuration with minimum cost, and with maximum overlap with ground truth. (b) Evaluation of the sampling framework. The y-axis shows the number of frames for which the true arm location was *not* found. Note the superior performance of the max-marginal distribution, especially including all our extensions (5), compared to using the marginal distribution with the best combination of our extensions.

5 Results

We evaluated our method against ground truth, and compared it to a simpler method using hand detection and tracking alone. This section reports and discusses the results.

Datasets: All evaluations were performed using a continuous sequence of 6,000 frames taken from BBC footage with challenging image conditions and a changing background (see Figure 1). Ground truth was manually labelled for 296 randomly chosen frames from this sequence. The corner of the image containing the signer was cropped and down-sampled to approximately 100×100 pixels. We concentrate on the more difficult case where the signer has sleeves with a similar colour to the torso. In the following, we refer to the arm on the left side of the image as the "left" arm.

Overlap measure: Quantitative evaluation was performed using an overlap measure defined as $o = \frac{T \bigcap M}{T \bigcup M}$, where T is a manual ground truth segmentation and M is the mask generated from an estimated configuration. We evaluate this overlap score separately for the left arm, the right arm and the hands. The overall overlap is defined as the mean over these three values. Note that this score takes occlusions into account i.e. the overlap is high only if the model and the true (not just the visible) area overlap.

We consider an overlap to be correct if it is ≥ 0.5 since this corresponds to a very good agreement with ground truth; overlaps between 0.2 and 0.5 are considered to be partially correct; overlaps below 0.2 are considered incorrect. An arm configuration is considered close to the configuration with greatest overlap if the difference in overlap is less than 0.1.

Evaluation of the complete cost function: Ideally, we would like to evaluate our complete cost function by exhaustive evaluation over the parameter space of both arms. Since this is computationally infeasible we illustrate the correlation between cost and overall overlap by fixing the left arm at the optimal position and evaluating over the right arm, see Figure 4(a) and quantitative results in Table 1 (columns C and CH). Our results show that the right arm can be found with very high accuracy. Detection of the left arm is also very good: 95.9% of 296 frames have an overlap ≥ 0.2. This is despite the changing background to the left of the signer. Adding HOG appearance features to the complete

	Left arm				Right arm				Hands
Accuracy	C	CH	CT	CHT	C	CH	CT	CHT	CHT
Overlap ≥ 0.2	95.9%	91.6%	100%	99.7%	99.7%	100%	100%	100%	100%
Overlap ≥ 0.5	83.8%	85.8%	84.1%	91.2%	96.3%	99.3%	97.3%	99.7%	95.6%
Overlap ≥ 0.6	59.5%	74.3%	64.2%	75.0%	82.4%	92.9%	84.1%	82.8%	82.8%

Table 1: Evaluation of 296 model fitting results for the left arm, right arm, and hands. The table shows the percentage of images with an overlap with ground truth above a given threshold. Ideally this number should always be 100%. Experiments were performed using colour cues (C), HOG cues (H), our distinctive frames tracking framework (T), and a combination thereof.

cost function did improve the number of highly accurate left and right arm positions, but at the expense of not finding the left arm in some frames.

Evaluation of the pictorial structure sampling framework: The sampling framework was evaluated by counting the number of images for which no sample was generated close to the true arm configuration. In total 296 images were used and 1,000 samples drawn per frame. As shown in Figure 4(b), using the max-marginal clearly outperforms the marginal distribution. Furthermore, our extensions lead to a decrease in the number of times the true arm configuration was *not* sampled from 22 to only 2 out of 296 images.

Evaluation of the distinctive frames temporal tracking approach: We evaluated the performance of our distinctive frames approach (see Table 1). Especially for the left arm, adding temporal information boosted the accuracy from 91.6% to 99.7%. Adding HOG features also improved the proportion of high overlaps significantly. Generally, the presented tracking results are extremely good for the left arm, the right arm and the hands.

We compared our results to a simpler method where (i) multiple hand candidates are detected for each frame by an AdaBoost detector and (ii) a Viterbi algorithm is used to assign detections to the left/right hand using a simple spatiotemporal prior. This simpler "hand only" method returned the wrong hand positions in 34 out of 296 frames (in contrast, our methods finds the true hand position in all frames). This is due to confusions between the left and right hand, the background containing hand-like shapes, and hands being "lost" when in front of the face. By solving the more difficult problem of finding the arms, not only does the hand detection accuracy increase, but also we extract information important for sign recognition such as the hand angles and the position of the elbows.

In the 6,000 frame signing sequence, 191 frames were identified to be distinctive. These 191 distinctive frames are distributed quite uniformly over the whole sequence, and hence losing track is not an issue. The identified arm position is incorrect in only one of these frames, due to the background having a sleeve-like colour.

We show the robustness of our approach on three videos[1] with different signers (see Figure 5). Our algorithm takes on average 2 minutes per frame on a 1.83 GHz machine.

6 Conclusions and Future Work

We proposed a generative model which can reliably find the arms and hands in sign language TV broadcasts with continuously changing backgrounds and challenging image conditions. Our algorithm requires minimal supervision, and works well on very long continuous signing sequences. This exceeds the state-of-the-art for continuous limb track-

[1] Available at `http://www.robots.ox.ac.uk/~vgg/research/sign_language/`

Figure 5: Extracts from 1 hour tracked sequences for 3 different signers.

ing. Possible extensions to the current model include the addition of a hand appearance term to the complete cost function, and automatic initialisation (no manual training).

Acknowledgements: We are grateful for financial support from the Engineering and Physical Sciences Research Council, Microsoft and the Royal Academy of Engineering.

References

[1] H. Cooper and R. Bowden. Large lexicon detection of sign language. *ICCV, Workshop HCI*, 2007.

[2] N. Dalal and B Triggs. Histogram of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[3] A. Farhadi, D. Forsyth, and R. White. Transfer learning in sign language. In *Proc. CVPR*, 2007.

[4] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient matching of pictorial structures. In *Proc. CVPR*, 2000.

[6] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computer*, 1973.

[7] A. Fossati, M. Dimitrijevic, V. Lepetit, and P. Fua. Bridging the gap between detection and tracking for 3D monocular video-based motion capture. In *Proc. CVPR*, 2007.

[8] T. Kadir, R. Bowden, E. J. Ong, and A. Zisserman. Minimal training, large lexicon, unconstrained sign language recognition. In *Proc. BMVC*, 2004.

[9] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Learning layered motion segmentations of video. *Proc. ICCV*, 2005.

[10] M.W. Lee and I. Cohen. A model-based approach for estimating human 3D poses in static images. *IEEE PAMI*, 2006.

[11] Z. Lin, L.S. Davis, D. Doermann, and D. DeMenthon. An interactive approach to pose-assisted and appearance-based segmentation of humans. In *ICCV, Workshop on Interactive Computer Vision*, 2007.

[12] R. Navaratnam, A. Thayananthan, P.H.S. Torr, and R. Cipolla. Hierarchical part-based human body pose estimation. In *Proc. BMVC*, 2005.

[13] E.J. Ong and R. Bowden. A boosted classifier tree for hand shape detection. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2004.

[14] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*. MIT Press, 2006.

[15] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Proc. CVPR*, 2005.

[16] L. Sigal and M.J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *Proc. CVPR*, 2006.

[17] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. *BMVC*, 2006.

[18] T. Starner, J. Weaver, and A. Pentland. Real-time American sign language recognition using desk- and wearable computer-based video. *IEEE PAMI*, 1998.

[19] D. Tran and D. Forsyth. Configuration estimates improve pedestrian finding. In *NIPS*, 2007.

[20] C. Yanover and Y. Weiss. Finding the M most probable configurations using loopy belief propagation. In *NIPS*, 2003.