Int J Comput Vis (2011) 95:180–197
DOI 10.1007/s11263-011-0480-9

Upper Body Detection and Tracking in Extended Signing Sequences

Patrick Buehler · Mark Everingham · Daniel P. Huttenlocher · Andrew Zisserman

Received: 8 November 2009 / Accepted: 24 June 2011 / Published online: 12 July 2011

Abstract The goal of this work is to detect and track the articulated pose of a human in signing videos of more than one hour in length. In particular we wish to accurately localise hands and arms, despite fast motion and a cluttered and changing background.

We cast the problem as inference in a generative model of the image, and propose a complete model which accounts for self-occlusion of the arms. 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) efficient sampling from a pictorial structure proposal distribution to obtain reasonable configurations; (ii) identifying a large number of frames where configurations can be correctly inferred, and exploiting temporal tracking elsewhere.

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

P. Buehler · A. Zisserman
Department of Engineering Science, University of Oxford, Oxford, UK

P. Buehler
e-mail: patrick@robots.ox.ac.uk

M. Everingham (✉)
School of Computing, University of Leeds, Leeds, UK
e-mail: M.Everingham@leeds.ac.uk

D.P. Huttenlocher
Computer Science Department, Cornell University, Cornell, USA

Keywords Person tracking · Human pose estimation · Sign language · Gesture recognition

1 Introduction

We investigate the task of articulated pose estimation and tracking of a person using sign language in long sequences of continuous video. Our work is motivated by a long term goal of automatic sign language recognition (Buehler et al. 2009), where extraction 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. As illustrated in Fig. 1, 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).

1.1 Related Work and Motivation

Previous approaches for hand localisation and tracking in signing video have concentrated on locating the hands by using their skin colour (Cooper and Bowden 2007; Farhadi et al. 2007; Starner et al. 1998) or by hand detectors based on "Viola & Jones" (Viola and Jones 2002) sliding-window classifiers (Kadir et al. 2004; Ong and Bowden 2004) using Haar-like image features and AdaBoost training. 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 which 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.

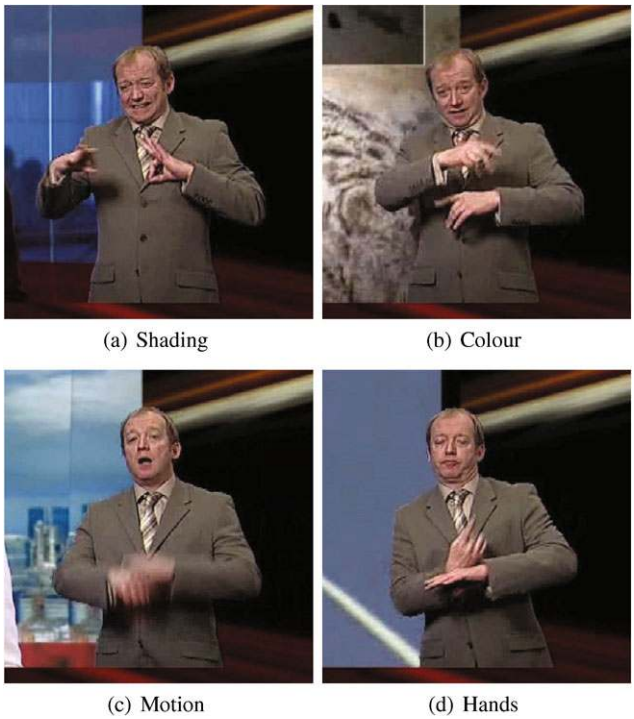


Fig. 1 Challenges for upper-body pose estimation. There are a number of image characteristics which render arm and hand detection ambiguous: (**a**) Shading and self-occlusions significantly change the appearance of the left arm. Note, here and in the rest of the paper, '*left*' refers to left in the image, rather than the person's 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 to left and right hand ambiguous. Note also that the background changes continuously

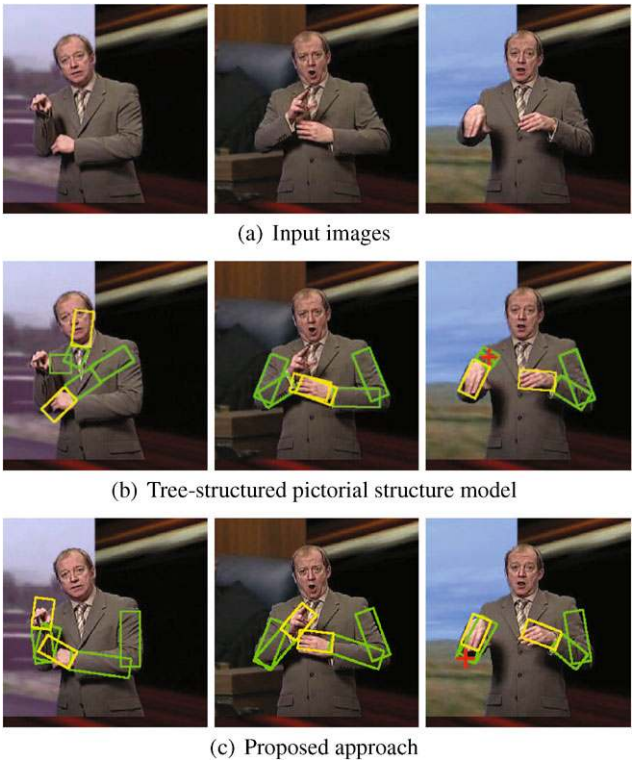


Fig. 2 Inherent drawbacks of tree-structured pictorial structure models. For the input images in (**a**), the corresponding images in (**b**) show arm configurations with high posterior probability under a tree-structured pictorial structure model; images in (**c**) show arm configurations estimated using the proposed method. (**b**, *left*) An example of *ignoring evidence*—the model only explains the foreground, i.e. pixels which are covered by the model. Each of the arm parts lie on the appropriate colour, even though this leaves skin pixels (the true hand) unexplained. (**b**, *middle*) An example of *over-counting of evidence*—each pixel can erroneously contribute evidence for multiple parts. In this example, the left and right hand of the estimated configuration are predicted to occupy the same image area. (**b**, *right*) If *occlusions* are not modelled correctly, the estimated upper and lower arms are very unlikely to lie at their true locations, which are occluded by the hand—note the estimated position of the left elbow marked with a cross

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 located) but in the end the constraint provided by the arms is worth the cost. Our approach is able to find the correct pose of the arms in hour-long continuous signing sequences, even though the background is complex (see Fig. 1), constantly changing (hence background subtraction would perform poorly), and can even contain other people. Furthermore, the arms move quickly, can be significantly foreshortened, do not follow simple motion patterns, and have a similar colour to the torso.

There has been much previous work on 2D human pose estimation, mainly using pictorial structures (Felzenszwalb and Huttenlocher 2000; Fischler and Elschlager 1973) based on tree-structured graphical models. Their great advantage is the low complexity of inference (Felzenszwalb and Huttenlocher 2005) and hence they have been used in numerous applications (Ramanan et al. 2005; Ramanan 2006; Sivic et al. 2006). While the run-time performance is very compelling, this approach has several limitations as a consequence of the independence assumptions implied by the use of a tree-structured model (see Fig. 2): (i) *ignoring evidence*—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*—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. Sigal and Black 2006); (iii) *no/poor modelling of occlusions*.

To overcome problem (ii) of over-counting, Lan and Huttenlocher (2005) augment the tree-like topology in order to capture correlations between pairs of parts not connected in the tree. Jiang (2009) finds the pose of a person in relatively uncluttered and static images by maximally covering extracted foreground silhouettes with an articulated model of the body.

Problem (iii) of un-modelled occlusions arises because tree-structured models limit the possibility of modelling occlusions to those between parent and child in the tree structure (Kumar et al. 2009) (though allowing individual parts to be occluded requires a doubling of the number of states for each part). One possible solution is to use multiple trees, but this also incurrs additional computational cost (Wang and Mori 2008). Alternatively, the restriction to trees can be removed. For example, Kumar et al. (2004) introduce a binary occlusion indicator for every part, and Sigal and Black (2006) approximate the global image likelihood using per-pixel binary variables which encode the occlusion relationships between parts. Both approaches sacrifice the computational efficiency and global optimality of inference in tree-structured graphs, and instead use variants of loopy and approximate belief propagation for inference.

Instead of tackling the main drawbacks of pictorial structures, other work has focused on improving appearance models for the individual parts. Gradient based cues have been used to design better part detectors, for example based on templates of Histogram of Oriented Gradients (this work), variants of shape context descriptors (Andriluka et al. 2009), or discriminatively learned Histogram of Oriented Gradients descriptors (Johnson and Everingham 2009). Eichner and Ferrari (2009) have improved the colour cue by exploiting relations between the appearance of different body parts.

Other approaches detect the pose of a person by iteratively assembling a person from low-level features, from part detections, or from locally optimal candidate configurations (Fleck et al. 1996; Micilotta et al. 2005; Siddiqui and Medioni 2007); or by learning a direct mapping from the image to the 3D pose (Agarwal and Triggs 2006).

1.2 Outline

We present a model which addresses the identified problems of (i) ignoring evidence, (ii) over-counting, and (iii) lack of occlusion modelling. The model includes an articulated part configuration for the foreground and also a model for the background. Every pixel in the image is generated using the part model or the background model. Self-occlusions are taken properly into account on a per-pixel basis and hence parts can be partially or fully occluded. In addition, the depth ordering of the two arms is modelled to indicate which arm is closer to the camera. Our model is similar in form to some previous work (Fossati et al. 2007; Kumar et al. 2005; Lee and Cohen 2006; Lin et al. 2007; Sigal and Black 2006) and described in detail in Sect. 2. Although we show that the pose with minimum cost correlates well with the true configuration of the upper body, it is too expensive to fit exhaustively. Consequently, we propose inference methods which avoid such search, but achieve acceptable results. Inspired by the work of Felzenszwalb and Huttenlocher (2005), we propose a sampling-based method for single frames where a pictorial structure is used as a proposal distribution (Sect. 3). Modifications to the sampling framework are introduced to generate a higher number of samples around the true arm configuration. Most notably this is achieved by sampling from the max-marginal rather than the marginal distribution. Temporal information is added by identifying "distinctive" frames for which the correct pose of the signer can be detected with high accuracy, and subsequently linking these frames by tracking pose configurations (Sect. 4). The advantage of this method is increased robustness and accuracy since temporal information helps to resolve otherwise ambiguous frames. Furthermore, execution time is reduced since the arm configuration in all non-distinctive frames can be found by tracking over time, which is faster than using a sampling-based approach. We evaluate our model and inference procedures on continuous signing footage taken from BBC broadcasts, using ground truth annotations. Quantitative results are reported in Sect. 5. Finally, Sect. 6 offers conclusions and directions for future work.

This submission is an extended version of our BMVC 2008 paper (Buehler et al. 2008). In addition to more detailed exposition and experiments, we present here revised methods for appearance modelling of the arms which reduces the requirement for training data, an improved procedure for segmentation of the head and torso, and more principled methods for the identification of distinctive frames.

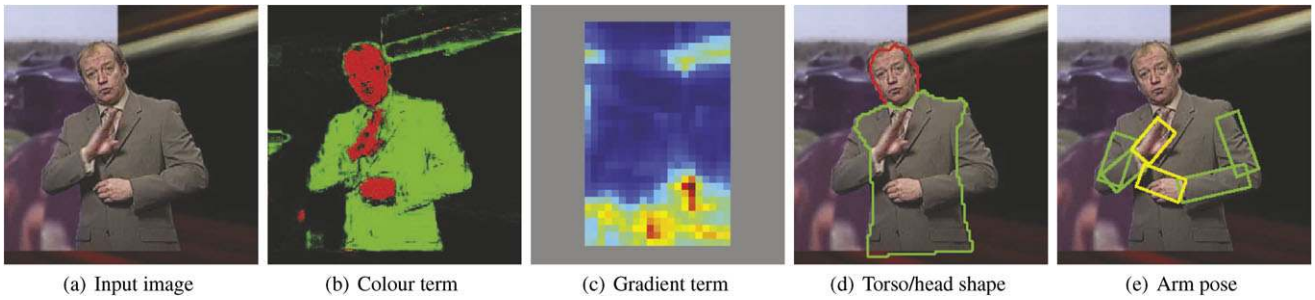
2 Generative Hand and Arm Model

This section describes our generative model which explains every pixel in the image and hence also takes into account the background as well as occlusions. We start by introducing our complete cost function which assigns a cost to a given configuration of the upper body. To reduce the complexity of modelling and inference, the pose estimation process is divided into two stages: First, the shape of the head and torso and the position of the shoulders are estimated (Sect. 2.3). Second, the configuration of both arms and hands are estimated as those with minimum cost given the head and torso segmentations. We describe the cost in this section, and then methods for obtaining the minimum cost efficiently in Sects. 3 and 4.

In the following, we refer to the arm on the left side of the image as the "left" arm, and respectively the arm on the right side of the image as the "right" arm.

2.1 Complete Cost Function

Formally, given a rectangular sub-image $\mathbf{I}$ that contains the upper body of the person and background, we wish to find



(a) Input image (b) Colour term (c) Gradient term (d) Torso/head shape (e) Arm pose

Fig. 3 (Color online) Overview of pose estimation process. Pose estimation for a given image (**a**) is performed using colour-based likelihoods (**b**) and likelihoods based on image gradients (**c**) (Sect. 2.2). The colour term in (**b**) is visualised by assigning the posterior probability for skin and torso to *red* and *green* colour channels respectively. The visualisation of the gradient term in (**c**) shows, for a given HOG template with fixed orientation and foreshortening, the likelihood at all locations in the image, where red indicates high likelihood. The example shown is for the right lower arm with orientation and foreshortening set to the ground truth values. Note the maximum is at the true centre of the right lower arm in the image. Using the colour term (**b**) the head and torso can be segmented (**d**) (Sect. 2.3). The arm pose (**e**) is then estimated using the predicted torso and head shape, and both colour and gradient terms (Sect. 3)

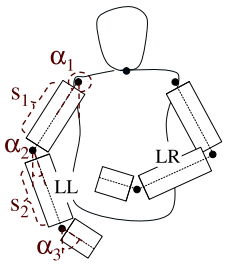


Fig. 4 Upper body model. The pose is specified by 11 parameters—5 for each arm and an additional binary parameter d indicating which arm is closer to the camera and hence visible in the case that the arms overlap. The shape of the head and torso and position of the shoulders are estimated in a pre-processing stage separate to estimation of the 2D arm configuration

the arm and hand configuration $\mathbf{L} = (\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n, d)$ which best explains the image, where $\{\mathbf{l}_i\}$ specifies the parts (limbs) and d 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 lower arms and the hands. The appearance (e.g. colour) and shape of the parts are learned from manual annotation of a small number of training images (see Sect. 2.2). The background is continuously varying, and largely unknown.

Every part $\mathbf{l}_i = (s_i, \alpha_i)$ is specified by two parameters: (i) an anisotropic scale factor s_i which represents the part's length relative to its width, in order to model foreshortening of the part due to out-of-plane rotation; (ii) orientation α_i representing in-plane rotation. Each part is connected to a single "parent" part such that the connections are in the form of a kinematic chain for the left and right arm respectively (see Fig. 4). While the upper and lower arm can each undergo foreshortening, the scale parameter for the two hands is fixed. By searching over the scale s_i for each part, our model can infer the correct arm configuration even in highly foreshortened cases, for example when the signer points towards the camera.

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

$$p(\mathbf{L}|\mathbf{I}) \propto p(\mathbf{L}) \prod_{i=1}^{N} p(\mathbf{c}_i|\lambda_i) \prod_{j\in\{LL,LR\}} p(\mathbf{h}_j|\mathbf{l}_j) \tag{1}$$

where N is the number of pixels in the input image, $\mathbf{c}_i$ is the colour of pixel i, and $\mathbf{h}_j$ is a HOG descriptor computed for limb j (see Sect. 2.2). The complete cost function is then defined as the negative logarithm of (1).

This cost function incorporates two appearance terms modelling the agreement between the image $\mathbf{I}$ and configuration $\mathbf{L}$. The first, $p(\mathbf{c}_i|\lambda_i)$, 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 depth ordering of the two arms is given by the binary variable d which specifies which arm is closer to the camera and hence visible in the case that the arms overlap. The "labelling" function $\Lambda(\mathbf{L}, i)$ is defined algorithmically essentially by rendering the model (Fig. 4) 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 Sect. 2.2). 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 upper body model.

The second appearance term, $p(\mathbf{h}_j|\mathbf{l}_j)$, models the likelihood of observed gradients in the image. This is based on HOG descriptors for the left and right lower arms (referred to in (1) as "LL" and "LR"). Both likelihood terms are described in more detail in Sect. 2.2.

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, and enforces the kinematic chain. The priors here are modelled as uniform in a pre-defined range, i.e. as simple constraints, in contrast to the spring-like quadratic costs between parts used in most previous work (Felzenszwalb and Huttenlocher 2000).

2.1.1 Complexity of Inference

There are 11 degrees of freedom in the model: 5 for each arm and 1 for the depth ordering. The state spaces of the arm parts are discretized into 12 scales and 36 orientations. The hand orientation is restricted to be within 50 degrees relative to the lower arm and discretized into 11 orientations. 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 Sect. 3 addresses this problem using a sampling approach.

2.2 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.

2.2.1 Colour Cue

The colour distribution for each of the body parts is modelled by a mixture of Gaussians using manually labelled data. Signed TV broadcasts are typically recorded in studio environments with controlled and constant lighting. Therefore, the foreground colour distribution can be learned from just a few training frames (see Sect. 2.4). Since the signer appears only in a corner of the employed TV footage, the background colour distribution is learned online from the remaining image area. This distribution is modelled by an RGB histogram and updated every frame to account for the changing background.

Having learned the colour distributions for each body part, every image pixel i with colour $\mathbf{c}_i$ is assigned a likelihood $p(\mathbf{c}_i|\lambda_i)$ for each label λ_i. Given a hypothesised configuration of parts $\mathbf{L}$, which implies a label λ_i for each pixel, the agreement with the image in terms of colour is evaluated by lookup of the colour likelihoods according to the corresponding pixel labels (1).

2.2.2 Histogram of Gradients Cue

In previous work the appearance of the different parts has typically been described using edge information at their boundaries. In our case these boundaries are often weak and unreliable 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 Fig. 1). We exploit both boundary and internal features to determine the position and configuration of a part using Histogram of Oriented Gradients (HOG) templates (Dalal and Triggs 2005; Tran and Forsyth 2007). The HOG descriptor represents the local distribution of image gradient orientation, over a grid of spatial bins. Two stages of contrast normalisation are applied to give invariance to photometric variation. Employing the HOG descriptor allows the appearance term to capture not only boundary edges but also internal edges. The agreement between the template of a given part with configuration $\mathbf{l}_i$ and the HOG descriptor of the image is evaluated using a distance function, described below.

A HOG template for each part, scale and orientation is learned from manually labelled images for which the pose of the upper body is given (see Fig. 5). The individual templates are computed as the mean over all training examples, and hence each template can be seen as a rotated and scaled version of every other template. Note that in Buehler et al. (2008) templates for a specific scale and orientation were learned only from training examples with similar scale and orientation. Although this allows modelling of effects such as gradient orientations due to the light source (and therefore independent of limb orientation), we found that in practise equal performance can be obtained by pooling examples across orientation and scale, requiring less training data.

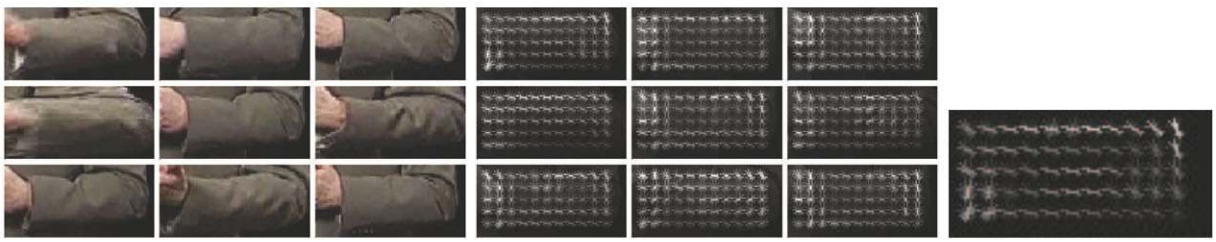
The likelihood function for the HOG appearance term, $p(\mathbf{h}_j|\mathbf{l}_j)$, is evaluated by computing the L2 distance between the image and the template and normalising to the range $[0, 1]$. In our experiments we compute the HOG appearance term for the left and right lower arms alone (i.e. not for the upper arms) since these are most often not occluded, and provide the strongest constraint on the wrist position. In contrast, occlusion of the upper arms, e.g. by the lower arm, is very common in signing and therefore we would not expect pose estimation to benefit greatly from adding upper arm cues.

2.2.3 Hand Labelling

As shown in Fig. 4, rectangles are used to model upper and lower arm parts when assessing the corresponding colour likelihood terms. While a rectangle is an adequate model for the arm parts, it is insufficient for the hands since they can assume many different shapes. 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). In effect this allows the hand shape to vary within the hypothesised oriented bounding box. For example, in Fig. 3(e),

(a) Training images for the right lower arm.

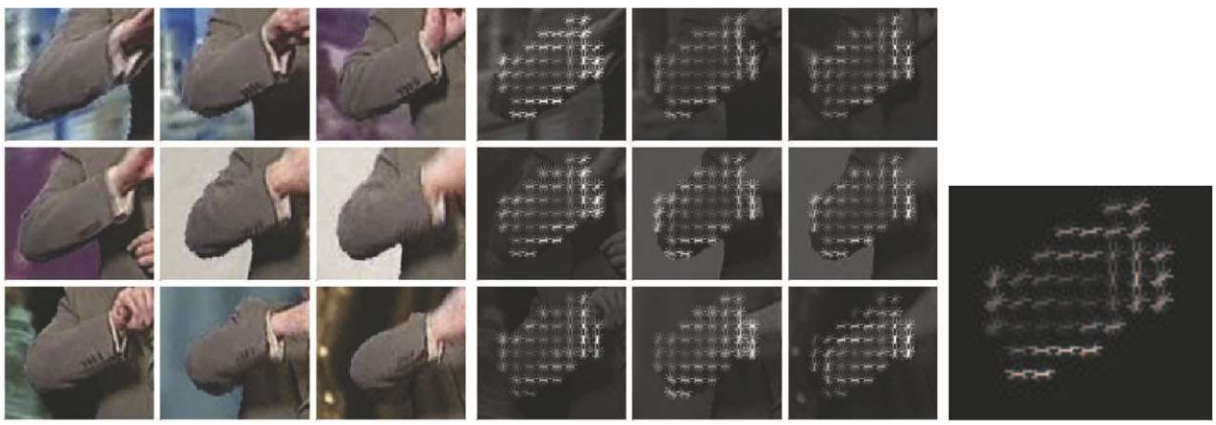
(b) Training images for left lower arm.



(c) Training patches extracted from images in (a) by cropping, rotation and scaling.

(d) HOG descriptors of patches in (c).

(e) Mean over HOG descriptors in (d).



(f) Close up of the training patches extracted from images in (b).

(g) HOG descriptors of patches in (f).

(h) Mean over HOG descriptors in (g)

Fig. 5 HOG template learning. The training data comprises images with the position of the lower arms manually annotated. To capture the inherent shape and internal features of the arm several templates are computed for each arm at different orientations and scales. This is illustrated for the right arm at horizontal orientation and maximum scale (i.e. no foreshortening) (**a**, **c**–**e**), and for the left arm at an orientation of 30 degrees and scale of 0.75 (i.e. foreshortening of 25%) (**b**, **f**–**h**). All training examples are used during training, independent of the target orientation and scale. Given a target orientation and scale, the lower arm is cropped and its co-ordinate system transformed (**c**, **f**) to match the target parameters. The HOG descriptor for every example is computed (**d**, **g**) and the mean over all examples used as the final template (**e**, **h**)

only pixels inside the two yellow hand rectangles and with a higher likelihood under the skin colour model than the torso or background colour models (see Fig. 3(b)) would be labelled as hand.

2.3 Head and Torso Segmentation

In a similar manner to previous work which has adopted a multiple stage approach to pose estimation (Ferrari et al. 2008; Navaratnam et al. 2005) in order to reduce ambiguity and computational expense, our method detects the shape of the head and torso and the position of the shoulders in a first stage before estimating the arm configuration. This two step approach is motivated by the restriction to frontal signing and the characteristic shape and colour of the torso which depends only weakly on the position of the arms (this is especially true if the sleeves and the torso share the same colour). This allows us to efficiently segment the head and torso with very high accuracy in a first stage, such that the explanation of the complete image by head, torso and arms in the second stage has high fidelity to the true segmentation. Note that this approach does not negatively influence (but rather helps) the subsequent arm pose estimation because of the precision of the estimated head and torso shape.

The benefits of this approach are: (i) a reduced search space by subsequently assuming known (and fixed) position of the head, torso and shoulders; (ii) a fast method described in Sect. 4 to identify frames where the pose can likely be estimated with high confidence based on counting the number of hand-like objects on the torso.

We approach the problem of segmenting the head and torso using multiple candidate shapes as templates (binary masks) and fitting these templates to the image using a simple two part pictorial structure model (see Fig. 6). Each part is specified by four degrees of freedom: 2D position, orientation and (isotropic) scale. The posterior distribution for this two part model is similar to (2) but here the appearance term uses only part-specific colour distributions, while the prior probability of a configuration enforces that the head is connected to the neck. For a given template, orientation and scale, the appearance term at every position is computed by convolution of the pixel-wise likelihoods under the torso colour model with the template. The final appearance term for the torso is then defined for each pixel individually as the maximum over all orientations and scales. Appearance terms for the head are computed in a corresponding manner, using the head shape templates and skin colour model.

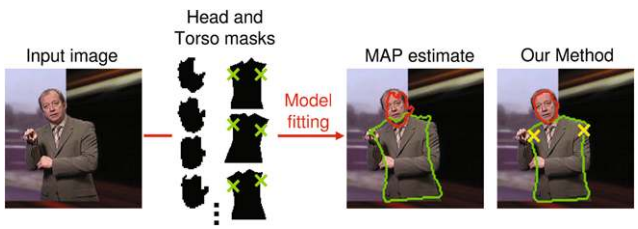


Fig. 6 Overview of head and torso segmentation. Candidate masks of the head and torso are fitted to the image using a pictorial structure model with two parts. The *maximum a posteriori* (MAP) segmentation is restricted to the head and torso shapes provided by the masks. In contrast, our method is based on a weighted nonlinear combination of the masks and hence achieves a more accurate segmentation. The position of the shoulders (crosses) is obtained by projecting the position of the shoulders from the torso masks into the image

Evidence from all the candidate shapes is combined to estimate the segmentation of the head (and similarly for the torso): (i) The posterior probability $p_H(x, y|\mathbf{I})$ of the head being at any given spatial position (x, y) in the image is computed. This requires marginalising over all possible torso configurations, which can be performed efficiently due to the restriction to tree-like topologies (see Sect. 3.2). In most cases p_H is unimodal, with a single peak centred on the MAP estimate. (ii) Sharpening is applied to amplify the mode(s): $p'_H(x, y|\mathbf{I}) = p_H(x, y|\mathbf{I})^\nu$ with $\nu > 1$. This is necessary since the posterior probability for the pictorial structure model is defined only up to a multiplicative constant. (iii) The sharpened distribution is then convolved with all head templates to obtain a score for each pixel: $s_H(x, y|\mathbf{I}) = \sum_{i=1}^{n_H} w_i\, p'_H(x, y|\mathbf{I}) * \mathbf{M}_i$. This score indicates if a pixel belongs to the background or to the head. The total number of head templates, including rotated and scaled variants, is referred to as n_H, and $\mathbf{M}_i$ denotes the binary mask of template i. The weight w_i is defined as the maximum response of template i over all (x, y) positions. The resulting value is taken to the power of η ($\eta > 1$) to effectively suppress head templates which do not describe the image well. The definition of the response is identical to the appearance term at the start of this section, and is computed by convolution of the pixel-wise likelihoods under the head colour model with the template mask $\mathbf{M}_i$. (iv) A threshold ζ is applied to $s_H(x, y|\mathbf{I})$ to classify each pixel as either head or background.

The segmentation of the torso is estimated in a corresponding fashion, marginalising over the head configuration. The position of the shoulders is similarly estimated as the weighted linear combination of the shoulder positions in each template. The parameters were estimated empirically and fixed to $\nu = 7$, $\eta = 10$ and $\zeta = 0.2 \max_{x,y} s_H(x, y|\mathbf{I})$. Note that the exact values of these parameters are not crucial to performance.

The approach described above differs from the original method reported in Buehler et al. (2008) in that we do not rely on a correct MAP estimate of the head and torso parts but instead use the (sharpened) marginal probability for each part (see step (i)). Note also that combining information from multiple templates by marginalisation allows the method to adapt the shape of the segmented body parts beyond the fixed set of training templates (see Fig. 6).

2.4 Learning the Model

Manually labelled data is required to learn the part-specific colour distributions, to build the head and torso model, and to create HOG templates. The colour distributions are learned from 5 frames, in which the visible area of each part has been segmented manually. For head and torso segmentation the shapes of 20 examples are provided manually. Note that the shapes of the head and the torso do not depend on possible occlusions from the hands and arms. HOG templates are learned from 39 images where the true arm configuration is 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 spaced over 0–180°, i.e. using unsigned gradients (Dalal and Triggs 2005).

3 Computationally Efficient Model Fitting

As will be demonstrated in Sect. 5, the minimum of the complete cost function in (1) correlates very well with the true arm configuration. However, the vast number of possible limb configurations makes exhaustive search for a global minimum of the complete cost function infeasible. In Sect. 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.

We will see in Sect. 4 that this sampling-based approach is used only to identify frames where the pose can be estimated with high confidence. In the following, we refer to these frames as *distinctive frames*. To reduce runtime complexity, the pose in each of the remaining frames is then estimated using a tracking-based approach initialised by the pose estimated for the distinctive frames.

3.1 Iterative Arm Fitting

The configuration of the arms is estimated in two passes. First, the best location of the left arm alone is found (with no right arm), then the best location of the right arm while keeping the left arm fixed at its best location. The process is then repeated with the ordering reversed, that is the right arm is fitted first. The minimum complete 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 the fitted 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 combine iterative arm fitting with a *stochastic* search for each arm, using an efficient sampling method (Felzenszwalb and Huttenlocher 2005) to propose likely candidate configurations. This reduces the complexity well below $O(N)$ by sampling 1,000 possible configurations for each arm. We also investigated inferring the configuration of both arms jointly. However, we found that an impractically high number of samples is necessary to obtain pose estimates that are of comparable accuracy.

Tree-structured pictorial structures are well suited for sampling-based inference since samples can be drawn efficiently from the corresponding distribution (Felzenszwalb and Huttenlocher 2005). However, as noted in the introduction this model has several shortcomings 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 complete generative model with the computational efficiency of tree-structured pictorial structure models.

The posterior distribution from which samples are drawn is given (Felzenszwalb and Huttenlocher 2005) by

$$p(\mathbf{L}|\mathbf{I}) \propto p(\mathbf{L}) \prod_{i=1}^{n} p(\mathbf{C}_i|\mathbf{l}_i) \tag{2}$$

where $\mathbf{L} = (\mathbf{l}_1, \ldots, \mathbf{l}_n)$ defines the configuration of each part and $\mathbf{C}_i$ refers to the pixels covered by part i. $p(\mathbf{L})$ is defined as in Sect. 2 and places plausible limits on the joint angles of the hands relative to the lower arms.

The appearance term $p(\mathbf{C}_i|\mathbf{l}_i)$ is composed of the product of pixel likelihoods within the rectangular body part region, using colour distributions modelled by mixtures of Gaussians, and edge and illumination cues added through HOG descriptors. Note that our appearance terms differ from previous work (Felzenszwalb and Huttenlocher 2005) in that we do not make use of a "centre-surround" filter, which assumes a body part is surrounded by a region of differing colour. As demonstrated in Fig. 7, this approach is not suitable for our domain, where the arms are often in front of the torso, since the arm and torso are typically similar in colour. A consequence is that the bottom-up appearance cues are more ambiguous, a difficulty overcome by modification of the sampling procedure (Sect. 3.3) and verification using the complete cost function (Sect. 2.1).

Sampling from (2) is facilitated by the restriction to tree-like topologies and can as a result be performed iteratively

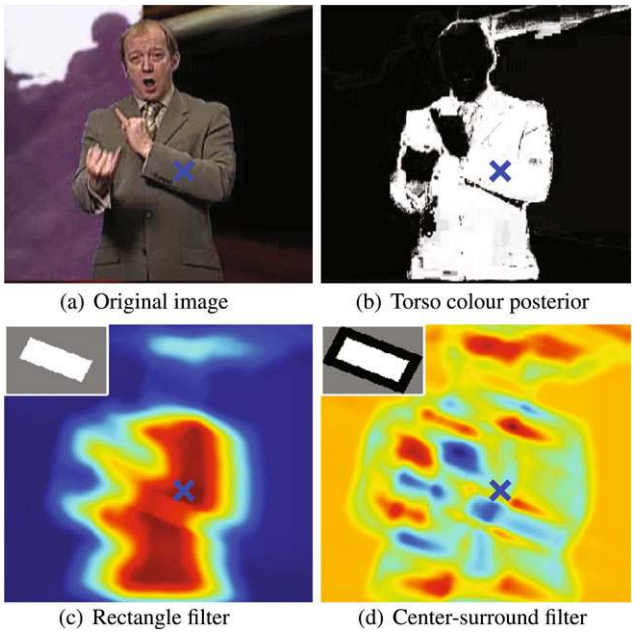


Fig. 7 (Color online) Influence of centre-surround filters on colour likelihood. For each pixel in image (**a**), the probability under the torso colour model is evaluated (**b**). The response of the right lower arm part assuming known orientation and foreshortening is computed using (**c**) a solid rectangle filter, or (**d**) a centre-surround filter. The filters are displayed in the upper left corner of (**c**) and (**d**), with the colours *black*, *grey*, and *white* corresponding to the values '−1', '0', and '1' respectively. *Red areas* in the heat plots indicate a high likelihood of the lower arm being at a certain position, while blue indicates a low likelihood. Note that the filter response around the true location of the right lower arm (indicated by a *blue cross*) is high only for the solid rectangle feature. This is due to the sleeves and the torso being of similar colour

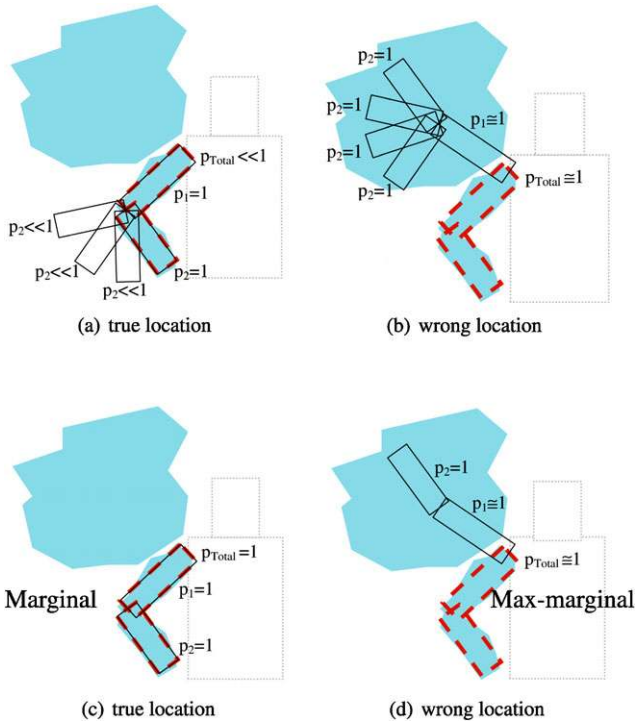


Fig. 8 (Color online) Sampling from max-marginal vs. marginal distribution. This example illustrates that drawing samples of the upper arm from the max-marginal distribution can be superior to using the marginal. Figures (**a**) and (**b**) show two cases where the upper arm rectangle is either placed 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 over the lower arm poses 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 over the lower arm poses is high due to a large area with arm-like colour 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 for the upper-arm, samples will be generated more often in the true arm location than using the marginal

(Felzenszwalb and Huttenlocher 2005). That is, the location of an arbitrary root node $\mathbf{l}_r$ can be sampled first by computing the marginal distribution $p(\mathbf{l}_r|\mathbf{I})$. Given $\mathbf{l}_r$, the location of all child parts can then be sampled recursively until all parts are instantiated. The marginal distribution for the root location is given as

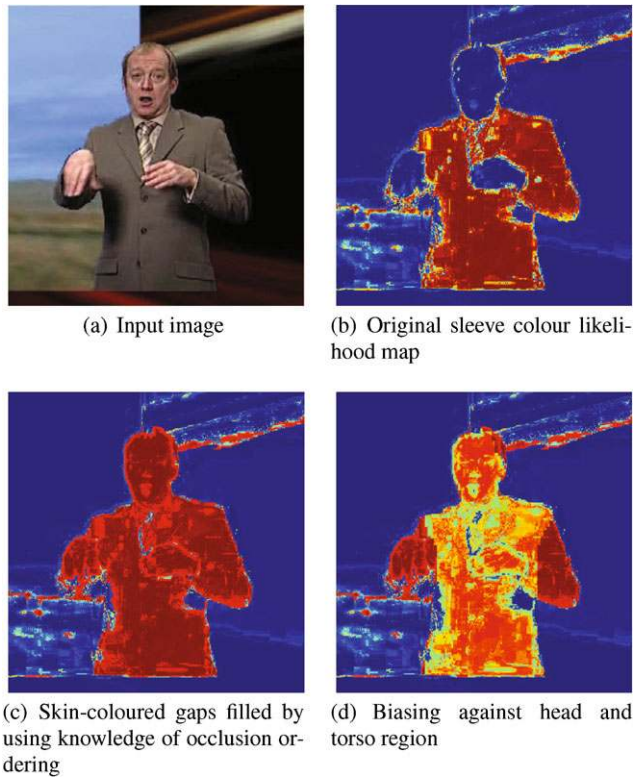
$$p(\mathbf{l}_r|\mathbf{I}) \propto \sum_{\mathbf{l}_1} \cdots \sum_{\mathbf{l}_{r-1}} \sum_{\mathbf{l}_{r+1}} \cdots \sum_{\mathbf{l}_n} \left(p(\mathbf{L}) \prod_{i=1}^{n} p(\mathbf{C}_i|\mathbf{l}_i) \right) \tag{3}$$

Computing this marginal directly as written above would take exponential time. By exploiting independence in the appearance terms $p(\mathbf{C}_i|\mathbf{l}_i)$ and independence between parts embodied in the tree-structured prior $p(\mathbf{L})$ a configuration can be sampled in time linear in the number and configurations of parts (Felzenszwalb and Huttenlocher 2005).

Samples can be drawn from (2) using the marginal distributions given in (3). However, we argue below that the use of max-marginals is better suited for this task, where the summation operation of standard marginalisation is replaced by maximisation:

$$p'(\mathbf{l}_r|\mathbf{I}) \propto \max_{\mathbf{l}_1} \cdots \max_{\mathbf{l}_{r-1}} \max_{\mathbf{l}_{r+1}} \cdots \max_{\mathbf{l}_n} \left(p(\mathbf{L}) \prod_{i=1}^{n} p(\mathbf{C}_i|\mathbf{l}_i) \right) \tag{4}$$

As demonstrated in Fig. 8, the intuition here is that a max-marginal sample of a "parent" part (e.g. the upper arm) is likely to be good if there is *any* configuration of the child (e.g. the lower arm) which has high probability in terms of prior and appearance. This is in contrast to a parent sample drawn from the marginal distribution which requires *all* compatible configurations of the child to be probable *on average*. As the figure shows, when using appearance terms which do not give a very sharp localised response, as in the case of filters with no centre-surround response, use of the



(a) Input image (b) Original sleeve colour likelihood map

(c) Skin-coloured gaps filled by using knowledge of occlusion ordering (d) Biasing against head and torso region

Fig. 9 Improving sampling efficiency by modified colour likelihood maps. The image in (**a**) shows a case where the left hand occludes most of the left upper and lower arm. The colour likelihood map for the sleeves is given in (**b**)—note that the probability of the left arm lying at its true position is very low. We allow for self-occlusions by modifying the likelihoods as shown in (**c**) where all gaps caused by skin-like colours are filled, making it more likely for the lower arm to lie at its true position. The example in (**d**) shows how a higher proportion of samples can be generated near the true arm position by lowering the likelihood that the torso and head pixels belong to the arms

marginal distribution places inappropriate weight on areas where the filters give flat response over a large area, and this bias is removed by use of the max-marginal distribution.

3.3 Domain-Specific Improvements in Sampling Efficiency

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 tree-structured pictorial structure sampler is insufficient; for example, given an image where a part is partially or completely occluded, the associated probability of sampling the true location for the part can be very low (see Fig. 9(b)). To increase the probability of sampling the true configuration, we propose the following modifications to the pictorial structure sampling framework, which exploit the restriction of our domain to frontal signing.

A. *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 make use of this ordering by modifying the colour likelihood term: The likelihood for a given pixel is redefined as the maximum likelihood for that pixel over the colour models corresponding to the part *and* all parts which could be occluding it (see Fig. 9(c)). Formally, we modify the colour likelihood $p(\mathbf{c}_i|\lambda_i)$ for the colour of a pixel i with label λ_i. Assuming a fixed occlusion ordering, then the part which corresponds to the label λ_i can only be occluded by the parts with labels $\Omega(\lambda_i)$, and hence the likelihood for pixel $\mathbf{c}_i$ is redefined as: $p'(\mathbf{c}_i|\lambda_i) = \max_{k\in\{\lambda_i\cup\Omega(\lambda_i)\}} p(\mathbf{c}_i|k)$.

   This modification increases the chance of an occluded part being correctly sampled, although at the expense of an on average higher proportion of bad samples. However, for our purpose, this is not problematic since sampling can be performed very efficiently.

B. *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 (Sect. 2.2) 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 (see Fig. 9(d)). Formally, this involves modifying $p(\mathbf{c}_i|\lambda_i)$ for all pixels that lie on the torso, $i \in \Psi$, and for all possible labels of a pixel i that share the torso colour model, $\lambda_i \in \Upsilon$, i.e. that have a colour similar to the torso. The likelihood is then redefined as: $p'(\mathbf{c}_i|\lambda_i) = p(\mathbf{c}_i|\lambda_i)^\kappa \ \forall i : i \in \Psi$ and $\forall\lambda_i : \lambda_i \in \Upsilon$. The parameter $\kappa > 1$ controls the strength of the bias; all reported experiments are performed with $\kappa = 3$.

   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. A similar approach is also used for parts which have the same colour as the head (in our case this is the hands) to avoid a high proportion of sampled hand positions within the head region.

C. *Sharpening instead of smoothing the probability distribution*. Felzenszwalb and Huttenlocher (2005) 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 Sect. 5) instead (that is to the take the distribution to the power of τ with $\tau > 1$, in contrast to smoothing where $\tau < 1$). This is mainly because the true arm configuration has a higher probability under the max-marginal than under the marginal

distribution (explained in Fig. 8) which is typically used to generate samples.

4 Tracking using Distinctive Frames

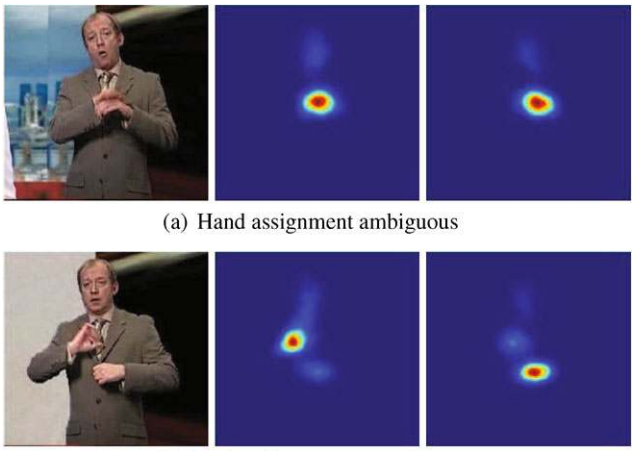
While we have concentrated thus far on pose estimation in isolated frames, for video sequences it is profitable to exploit the temporal coherence between frames. Especially for ambiguous poses, which may introduce multiple modes in the complete cost function, the use of temporal coherence can significantly improve the accuracy of pose estimation. We propose a method based on detection of "distinctive" frames, where the pose is unambiguous, and tracking between pairs of such frames.

We show that distinctive frames can be detected not more than a few seconds apart, and hence only a small number of frames are affected if the track is lost. This approach owes some inspiration to the work of Ramanan et al. (2005), 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.

Our method for identifying distinctive frames works by analysing features of the pictorial structure proposal distribution (2). We observe that most cases where the true pose is not identified are due to confusion between the left and right hands, or (less frequently) due to the background containing limb-like structures. This motivates finding unambiguous "distinctive" frames where (i) the hands are on the body and (ii) there is no confusion between the left/right assignment. As demonstrated in Fig. 10, such unambiguous frames can be identified by analysing the posterior distributions of the left and right hands.

To this end we employ a simple approach to label a given frame as distinctive as follows: (i) if two skin-coloured areas are present on the body, (ii) compute the posterior distributions of the left and right hands (Fig. 10, middle and right column), and (iii) evaluate a distinctiveness measure (described in the next paragraph). Finally, (iv) mark the frame as distinctive if this value is above a threshold.

We compare three different distinctiveness measures: entropy, mutual information, and dot-product. Section 5.6 reports a quantitative evaluation. Given the posterior distributions over the left and right hand position $p(\mathbf{x}_L)$ and $p(\mathbf{x}_R)$, the distinctiveness measures are defined as (i) entropy: $-(E(\mathbf{x}_L) + E(\mathbf{x}_R))$; (ii) mutual information: $-I(\mathbf{x}_L; \mathbf{x}_T)$; (iii) dot-product: $-\sum_{\mathbf{x}} p(\mathbf{x}_L)p(\mathbf{x}_R)$. The entropy measure evaluates the individual "peakiness" of the distributions for each hand. In contrast, the mutual information and dot-product measures evaluate the *dissimilarity* between the left and right hand posteriors.



(a) Hand assignment ambiguous

(b) Hand assignment unambiguous

Fig. 10 Identification of distinctive frames. For two frames (*left column*), the posterior of the left hand (*middle column*) and the right hand (*right column*) is computed from the pictorial structure proposal distribution (2). In (**a**), the similar spatial locations of the modes in the posterior indicate that the assignment to left and right hands is ambiguous, while in (**b**) the distinct modes indicate an unambiguous case. We identify unambiguous cases by a distinctiveness measure on these posteriors for each hand

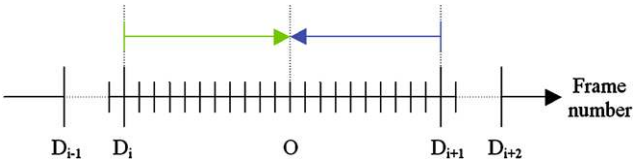


Fig. 11 Tracking using distinctive frames. First, distinctive frames D are identified where the pose can be estimated with high confidence. Poses are propagated from distinctive frames by tracking forward/backward in time (arrows). For distinctive frames D_i and D_{i+1} which are N frames apart, only $N/2$ frames need to be tracked from either end

Note that the methods proposed here differ to that used in Buehler et al. (2008), in that distinctive frames are identified using only the pictorial structure proposal distribution rather than the complete cost function. As a result, distinctive frames can be identified faster, but with similar accuracy (see Sect. 5.6).

4.1 Tracking Between Distinctive Frames

Detection of distinctive frames typically yields around one frame per second of video for which the pose can accurately be estimated with high probability. 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 (see Fig. 11).

Temporal tracking is realised by adding a tracking term $p(\mathbf{L}|\mathbf{L}') = \prod_{k=1}^{n} p(\mathbf{l}_k|\mathbf{l}'_k)$ to the complete cost function in (1) where $\mathbf{L}' = (d', \mathbf{l}'_1, \mathbf{l}'_2, \ldots, \mathbf{l}'_n)$ refers to the part configurations in the preceding frame. The conditional probability

$p(\mathbf{l}_k|\mathbf{l}'_k)$ is large if $\mathbf{l}_k$ and $\mathbf{l}'_k$ are similar and close 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 $p(\mathbf{l}_k|\mathbf{l}'_k) \propto p(\mathbf{l}_k - \mathbf{l}'_k)$ 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. Note that the motion model uses only first order (velocity) information; in contrast to specific actions such as walking or golf swings studied in previous work on pose estimation, the arm motion in sign language is much less predictable hence more complex models are not applicable.

Tracking the arms from frame to frame is greedy in the sense that we maintain only a point estimate of the pose rather than the full distribution over pose. This can potentially result in losing track by propagating incorrect poses. However, in practise this is not a problem since two distinctive frames are seldom more than a few seconds apart. The sustained human tracking method of Sheikh et al. (2008) also demonstrates the success, and gives a fuller Kalman filter treatment, of combining first order motion models with a distribution over pose detection.

5 Results

In this section we evaluate our method against ground truth, and compare it to a method which employs detection and tracking of the hands alone.

5.1 Datasets

All evaluations were performed using a continuous sequence of 6,000 frames taken from BBC footage[1] with challenging image conditions and a changing background (see Fig. 1). The corner of the image containing the signer was cropped and down-sampled to 100 × 100 pixels. We concentrate on the more difficult case where the signer has sleeves with a similar colour to the torso—when the signer wears short sleeves identification of the hand shape is difficult, but estimating the pose of the arms is considerably simplified.

Ground truth was manually labelled for 296 randomly chosen frames from this sequence. As shown in Fig. 14, each image was manually segmented to give masks for torso, head, upper and lower arms, and hands.

5.2 Overlap Measure

Quantitative evaluation was performed using an overlap measure defined as $o(T, M) = \frac{T \cap M}{T \cup M}$, where T is the ground truth segmentation and M the mask generated from an estimated pose (see Fig. 12). We evaluate the overlap separately for the left arm, the right arm and the hands. The overall overlap is then defined as the mean over the overlap for each body part. Note that this measure takes occlusions into account i.e. the overlap is high only if the model and the true (not just the visible) area overlap.

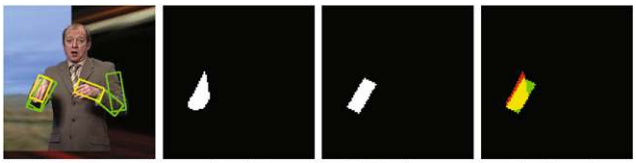


Fig. 12 Overlap evaluation measure. The estimated pose is shown in (**a**); (**b**)–(**d**) illustrate the overlap measure for the left upper arm. The overlap between ground truth (**b**) and estimated segmentation (**c**) is defined as the ratio of the intersection over the union (**d**). In this example, the overlap is 0.63

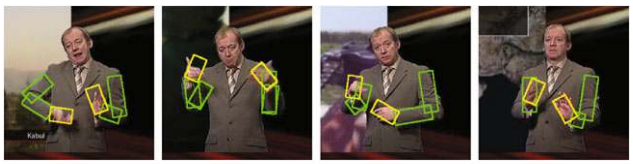


Fig. 13 Qualitative accuracy as a function of overlap measure. (**a**)–(**d**) show estimated poses for a range of overlap measures o. An overlap of $o = 1$ implies perfect segmentation of the image into the left arm, right arm and the hands. Note that the pose is qualitatively correct for overlap measures exceeding around 0.2

We consider an overlap to be correct if it is ≥ 0.5, which corresponds to a very good agreement with ground truth; overlaps between 0.2 and 0.5 are considered to be partially correct; and overlaps below 0.2 are considered incorrect. Furthermore, we define the true arm configuration as the one with highest overlap score, and consider an arm configuration as close to the true configuration if their difference in overlap is less than 0.1 (see Fig. 13 for examples).

5.3 Evaluation of the Complete Cost Function

As noted in Sect. 2 our approach uses a "complete" cost function which explains all pixels of the image (both signer and background). We first evaluate the effectiveness of the cost function i.e. the correspondence between an accurate estimated pose and low cost. Ideally, we would like to evaluate this 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 right arm at the optimal position and evaluating over the left arm. Figure 14 demonstrates the relationship between cost and overlap with ground truth. Note the good

[1] Images and ground truth available at: http://www.robots.ox.ac.uk/~vgg/data/sign_language/index.html.

Table 1 Quantitative evaluation. Pose estimation accuracy is reported for 296 ground truth frames, in terms of mean overlap for both arms and hands. The table shows the percentage of images with an overlap with ground truth above a given threshold. Experiments were performed using colour cues (C), HOG cues (H), the distinctive frames tracking framework (T), and combinations thereof

Accuracy	Left arm				Right arm				Hands
	C	CH	CT	CHT	C	CH	CT	CHT	CHT
Overlap ≥0.2	99.4%	98.3%	98.3%	98.3%	100.0%	100.0%	100.0%	100.0%	100.0%
Overlap ≥0.5	83.4%	90.1%	81.1%	86.8%	97.6%	98.8%	98.7%	99.0%	94.3%
Overlap ≥0.6	56.7%	70.3%	61.7%	73.9%	78.1%	88.5%	79.9%	81.3%	83.4%

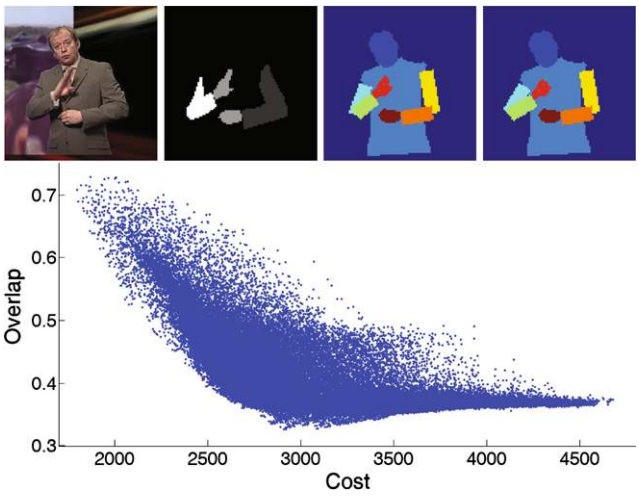


Fig. 14 Evaluation of the complete cost function. The *top row* shows, from left to right: input image, manual ground truth, pose having minimum cost, and pose having maximum overlap with ground truth. The *plot below* shows the value of the complete cost function vs. overlap with ground truth, calculated by fixing the right arm to the configuration with maximum overlap and evaluating all possible configurations for the left arm, including the depth order of the arms. Note that for low costs there is good correlation between cost and overlap

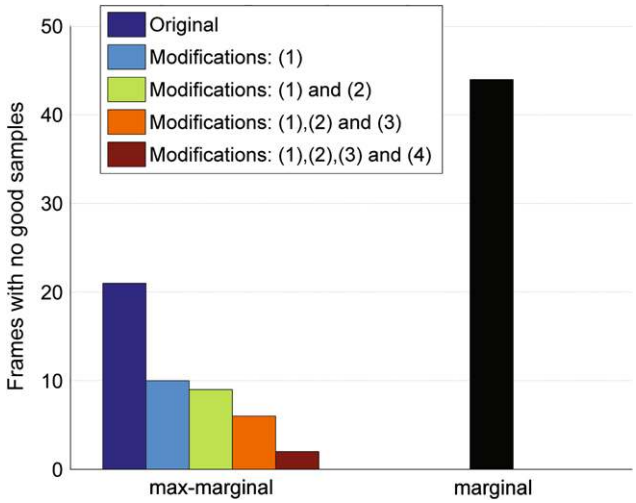


Fig. 15 Evaluation of sampling schemes. The y-axis shows the number of frames for which the true arm location is *not* found. "Original" refers to samples from the max-marginal without any modifications. The proposed modifications in Sect. 3.3 significantly reduce the number of errors. These are: (1) sharpening the proposal distribution, (2) adding HOG cues, (3) adding knowledge of the occlusion ordering, and (4) sampling less often within the head and the torso

correlation between high overlap measure and low cost, illustrating that the arm configurations with low cost coincide with the true position of the arms in the image.

Table 1 shows quantitative results for the 296 frames with ground truth annotation. Columns C and CH compare results using colour cues alone (C) and colour cues plus HOG descriptors for the lower arms (CH). The results show that the right arm can be found correctly in all frames. Detection of the left arm is also very reliable: 99.4% of 296 frames have an overlap ≥0.2. This is despite the changing background to the left of the signer. Adding HOG features to the complete cost function (column CH) substantially improves the number of frames for which the estimated pose is highly accurate (overlap ≥0.5). However, for the left arm the use of HOG features causes inaccurate localisation for some frames, where the appearance of the lower arm differs significantly from the available training examples.

5.4 Evaluation of Sampling Framework

The pictorial structure 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 Fig. 15, using the max-marginal clearly outperforms the marginal distribution. Furthermore, the proposed extensions (Sect. 3.3) 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.

The superior performance of the max-marginal is also replicated if the rectangle filters are replaced with centre-surround filters (as shown in Fig. 7). In the centre-surround case no samples were generated close to the true configuration in 30 images using the max-marginal, and in 99 using the marginal. Note that both these results are inferior to those using the rectangle filters; this can be attributed to the incorrect assumption made in the centre-surround filter that

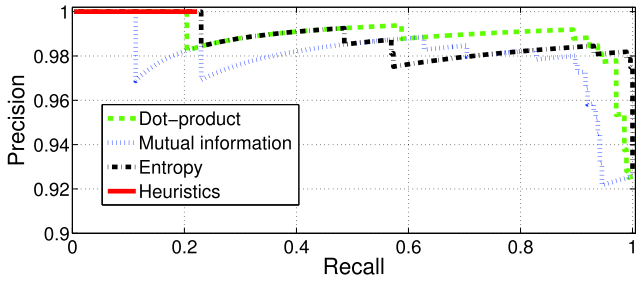


Fig. 16 (Color online) Identification of distinctive frames. The effectiveness of four methods for distinctive frame detection is evaluated in terms of precision/recall, using 296 frames with ground truth pose. We speak of a positive if a frame is classified as distinctive, and true positive if, in addition, the estimated pose is correct (i.e. overlap ≥ 0.2). Consequently, precision is measured as the proportion of frames classified as distinctive and with correctly identified arm pose, and recall is the proportion of frames classified as distinctive. *Curves* correspond to the four measures proposed to analyse the posterior distributions of the left and right hand (see Sect. 4): dot product (*green dashed curve*), mutual information (*blue dotted curve*), and entropy (*dash dotted curve*). For comparison, the accuracy of our original method to identify distinctive frames (Buehler et al. 2008) based on heuristics is shown by the red solid curve. Note that the y-axis starts at a precision of 90%

a limb is surrounded by background of differing colour—for the case of sign language, where limbs are often in front of the body, this assumption is often violated.

5.5 Evaluation of Distinctive Frame Detection

We evaluate the three proposed methods for detecting distinctive frames (Sect. 4) and compare to our original method proposed in Buehler et al. (2008). Our original method is based on a heuristic to count the number of modes in the complete model; with this method, out of the 296 frames used for evaluation, 61 were selected as distinctive. The estimated arm poses in these 61 frames are all correct. Figure 16 shows a comparison of the proposed distinctiveness measures with our original method. We adopt a precision/recall protocol, measuring precision as the proportion of frames which are classified as distinctive and for which the pose is correctly estimated (overlap ≥ 0.2). At low recalls (up to 0.2), our original method performs equally well to the distinctiveness measures from Sect. 4 based on the dot product or entropy. However the recall of our original method is low, classifying only around 20% of the frames as distinctive. Using the proposed dot product measure, up to 90% of the frames can be classified as distinctive, with a corresponding precision of 99%.

In the 6,000 frame signing sequence (see Sect. 5.1), 191 frames were classified as distinctive. Note that, to save execution time, once a distinctive frame is found, the following 10 frames are not considered as candidates for further distinctive frames. These distinctive frames are distributed quite uniformly over the whole sequence such that tracking seldom has to be performed for more than 1–2 seconds and hence losing track is not an issue. The identified arm position is incorrect in only one of these 191 frames, due to the background having a sleeve-like colour.

Table 2 Joint localisation error. Statistics of the Euclidean distance between true and estimated 2D positions are shown for the wrist, elbow and shoulder joints. Results for the wrists and elbows are qualitatively highly accurate (mean distance less than 3 pixels); for comparison the distance between the signer's eyes is around 7.5 pixels

	Wrist		Elbow		Shoulder	
	Left	Right	Left	Right	Left	Right
Mean	2.5	2.8	3.3	3.1	7.6	4.7
Median	2.1	2.7	3.0	2.8	7.4	4.8
Min	0.2	0.2	0.3	0.0	3.0	0.8
Max	12.3	7.2	12.2	9.3	12.5	9.5
Std.	1.7	1.3	1.9	1.7	1.9	1.6

5.6 Evaluation of Tracking Using Distinctive Frames

While results obtained on a per-frame basis (Sect. 5.3) are already highly accurate, we show that further improvements can be obtained by incorporating temporal information. Columns CT and CHT in Table 1 report results using our distinctive frames approach to identify frames for which the arm pose can be estimated with high confidence, and subsequently tracking between them (Sect. 4). Especially for the left arm, adding temporal information improves the accuracy for highly accurate poses from 70.3% to 73.9%.

Adding HOG features also improved the proportion of high overlaps significantly. Overall the tracking results are essentially qualitatively accurate ($o \geq 0.2$) for both arms and hands in all frames.

In addition to the proposed overlap measure, we present quantitative results in terms of error in estimated joint positions, as has been used in some previous work on 3D pose estimation. Table 2 reports statistics of the Euclidean distance between true and estimated 2D joint positions for the wrist, elbow and shoulder joints. The mean and median errors for the wrists and elbows are around 3 pixels. This is qualitatively highly accurate—for comparison, the distance between the eyes of the signer is around 7.5 pixels. Localisation of the shoulder joint is less accurate (up to 8 pixels), although this can be hard to localise visually, and is less important for interpretation of the signer's actions.

5.7 Comparison to Baseline Tracking Based on Hand Detection

We have noted that our proposal to localise the signer's hands using full upper-body pose estimation introduces complexity, but is necessary to overcome challenges such

as correct left/right hand assignment, or coping with background clutter including other hands. We compare our results to a method using a more conventional approach of hand detection and tracking to validate the effectiveness of estimating the full upper-body pose.

The hand detection-based method operates as follows: first, candidate hands are detected in each frame using a hand detector. The detector uses the well-known method of Viola and Jones (2002) (boosted decision stumps on Haar-like features), but employs two feature channels: (i) image intensity; (ii) a skin colour channel, representing the posterior probability that a pixel is skin-coloured. Non-maximum suppression is applied and all combinations of detections are enumerated as (a) left/right hand; (b) right/left hand or (c) two hands sharing a detection. This gives candidate interpretations of the detections for each frame, including the possibility that one hand occludes the other. Each interpretation is assigned a likelihood according to the classifier confidence for the corresponding detections (which can be considered an approximate log-likelihood ratio). Two additional likelihoods are defined: (i) a spatial prior on the position of left/right hands. The distribution over position for each hand is modelled as a single Gaussian with full covariance. This captures the weak prior that the left hand tends to appear to the left of the signer's torso, and vice-versa; (ii) a motion model over hand positions for consecutive frames. In this case a zero-mean isotropic Gaussian is assumed for the difference in position of each hand—as noted such simple motion models (here "zero velocity") prove more effective than more complex models for sign language, since the hand motion is constantly changing.

Spatial, motion and appearance (detection) likelihoods are combined to give an overall likelihood for each interpretation, given the interpretation of the previous frame. Consistent interpretations for each frame of the video are then selected by applying the Viterbi algorithm to maximise the joint likelihood of all frames simultaneously—this is tractable because of the Markov assumption in the motion model. Compared to particle filtering approaches to tracking, the proposed method gives a globally optimal interpretation, subject to the finite number of initial hand candidates (five were used) selected for each frame. Similar approaches have been applied for effective interactive point tracking in video, e.g. Buchanan and Fitzgibbon (2006).

The detection-based method predicted incorrect hand positions in 34 out of 296 frames (11%); in contrast, our upper-body method finds the true hand position in all frames. As shown in Fig. 17, errors are 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. Of these, incorrect left/right hand assignment is most prevalent—the weak priors on global hand position and hand motion are insufficient to obtain correct assignments.

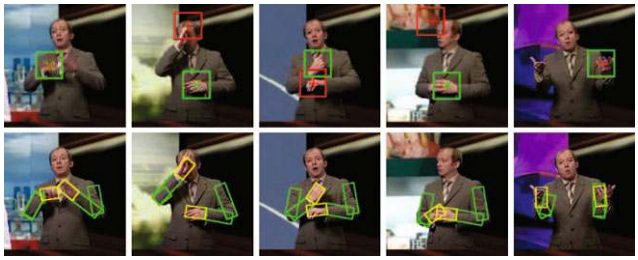


Fig. 17 Comparison of upper-body vs. hand-only tracking. The columns show failure modes of hand-only detection and tracking—the top image shows the hand positions estimated by the hand-only method (Sect. 5.7), and the bottom image shows the pose estimated by our method. Hand detection is challenging in these images due to (left to right): motion blur, hand in front of face, proximity of the hands, hand-like objects in the background, shadows

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 orientation and the position of the elbows.

5.8 Evaluation on Hour-Long Sequences

We have evaluated the robustness of our proposed approach on three hour-long video sequences with different signers. Videos of the results are available on the web.[2] Figure 18 presents a summary of the results, showing the estimated hand and arm configurations for frames taken at equal intervals throughout the sequences. In only 1 out of the 72 frames shown is the true arm configuration not found. Note that we are able to achieve good results even though the edge between sleeve and torso can be very weak (see Fig. 18, bottom signer).

5.9 Computational Expense

The run-time of the proposed method has two components: First, distinctive frames are identified which involves sampling of 1,000 possible arm configurations and evaluating each sample using our model. Second, tracking is performed between distinctive frames which tests on average 50,000 possible arm configurations per frame (this number was chosen conservatively and can be reduced by an order of magnitude with little influence on the overall accuracy).

The algorithm takes on average 100 seconds per frame on a 1.83 GHz machine implemented mostly in Matlab: it takes 20 seconds to segment the head and the torso, 90 seconds to identify a frame as distinctive and 60 seconds for tracking in between distinctive frames. Identification of distinctive frames is accelerated to an average of 20 seconds per frame by considering only frames where in a screening-step two skin-coloured areas are detected on the torso and

[2] http://www.robots.ox.ac.uk/~vgg/research/sign_language/index.html.

Fig. 18 Sample of results on hour-long sequences. The estimated pose is shown for uniformly spaced frames in three hour-long sequences with different signers. For all but one of the 72 frames shown (*row 6*, *column 1*) the estimated pose is qualitatively highly accurate

by skipping 10 consecutive frames once a distinctive frame is identified.

6 Conclusions and Future Work

We have 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. The model is a combination of a quite 'tight' (accurate) model of the foreground human together with a 'loose' model of the background. Our algorithm requires minimal supervision, and works well on very long continuous signing sequences. This exceeds the state-of-the-art for continuous limb tracking. Possible extensions to the current model include the addition of a more descriptive hand appearance term, and automatic initialisation (no manual training).

Acknowledgements We are grateful for financial support from the Engineering and Physical Sciences Research Council, Microsoft, the Royal Academy of Engineering, and ERC grant VisRec No. 228180.

References

Agarwal, A., & Triggs, B. (2006). Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(1), 44–58.

Andriluka, M., Roth, S., & Schiele, B. (2009). Pictorial structures revisited: people detection and articulated pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Buchanan, A. M., & Fitzgibbon, A. W. (2006). Interactive feature tracking using k-d trees and dynamic programming. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (Vol. 1, pp. 626–633).

Buehler, P., Everingham, M., Huttenlocher, D. P., & Zisserman, A. (2008). Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proceedings of the British machine vision conference*.

Buehler, P., Everingham, M., & Zisserman, A. (2009). Learning sign language by watching TV (using weakly aligned subtitles). In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Cooper, H., & Bowden, R. (2007). Large lexicon detection of sign language. In *ICCV, workshop human computer interaction* (Vol. 4796, pp. 88–97).

Dalal, N., & Triggs, B. (2005). Histogram of oriented gradients for human detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 886–893).

Eichner, M., & Ferrari, V. (2009). Better appearance models for pictorial structures. In *Proceedings of the British machine vision conference*.

Farhadi, A., Forsyth, D., & White, R. (2007). Transfer learning in sign language. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–8).

Felzenszwalb, P., & Huttenlocher, D. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision 61*(1), 55–79.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2000). Efficient matching of pictorial structures. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 2066–2073).

Ferrari, V., Marin-Jimenez, M., & Zisserman, A. (2008). Progressive search space reduction for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Fischler, M., & Elschlager, R. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computer*, *c-22*(1), 67–92.

Fleck, M. M., Forsyth, D. A., & Bregler, C. (1996). Finding naked people. In *Lecture notes in computer science: Vol. 1065. Proceedings of the European conference on computer vision* (pp. 591–602). Berlin: Springer.

Fossati, A., Dimitrijevic, M., Lepetit, V., & Fua, P. (2007). Bridging the gap between detection and tracking for 3D monocular video-based motion capture. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–8).

Jiang, H. (2009). Human pose estimation using consistent max-covering. In *Proceedings of the international conference on computer vision*.

Johnson, S., & Everingham, M. (2009). Combining discriminative appearance and segmentation cues for articulated human pose estimation. In *IEEE international workshop on machine learning for vision-based motion analysis*.

Kadir, T., Bowden, R., Ong, E. J., & Zisserman, A. (2004). Minimal training, large lexicon, unconstrained sign language recognition. In *Proceedings of the British machine vision conference*.

Kumar, M. P., Torr, P. H. S., & Zisserman, A. (2004). Extending pictorial structures for object recognition. In *Proceedings of the British machine vision conference*.

Kumar, M. P., Torr, P. H. S., & Zisserman, A. (2005). Learning layered motion segmentations of video. In *Proceedings of the international conference on computer vision*.

Kumar, M. P., Torr, P. H. S., & Zisserman, A. (2009). Efficient discriminative learning of parts-based models. In *Proceedings of the international conference on computer vision*.

Lan, X., & Huttenlocher, D. (2005). Beyond trees: common-factor models for 2D human pose recovery. In *Proceedings of the international conference on computer vision: Vol. 1*.

Lee, M., & Cohen, I. (2006). A model-based approach for estimating human 3D poses in static images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(6), 905–916.

Lin, Z., Davis, L., Doermann, D., & DeMenthon, D. (2007). An interactive approach to pose-assisted and appearance-based segmentation of humans. In *ICCV, workshop on interactive computer vision*.

Micilotta, A., Ong, E., & Bowden, R. (2005). Real-time upper body 3D pose estimation from a single uncalibrated camera.

Navaratnam, R., Thayananthan, A., Torr, P., & Cipolla, R. (2005). Hierarchical part-based human body pose estimation. In *Proceedings of the British machine vision conference* (pp. 479–488).

Ong, E., & Bowden, R. (2004). A boosted classifier tree for hand shape detection. In *Proceedings of the international conference on automatic face and gesture recognition* (pp. 889–894).

Ramanan, D. (2006). Learning to parse images of articulated bodies. In *Advances in neural information processing systems*. Cambridge: MIT Press.

Ramanan, D., Forsyth, D. A., & Zisserman, A. (2005). Strike a pose: tracking people by finding stylized poses. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (Vol. 1, pp. 271–278).

Sheikh, Y., Datta, A., & Kanade, T. (2008). On the sustained tracking of human motion. In *Proceedings of the international conference on automatic face and gesture recognition*.

Siddiqui, M., & Medioni, G. (2007). Efficient upper body pose estimation from a single image or a sequence. In *Human motion, lecture notes in computer science: Vol. 4814*.

Sigal, L., & Black, M. (2006). Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 2041–2048).

Sivic, J. Zitnick, C. L., & Szeliski, R. (2006). Finding people in repeated shots of the same scene. In *Proceedings of the British machine vision conference*.

Starner, T., Weaver, J., & Pentland, A. (1998). Real-time American sign language recognition using desk- and wearable computer-based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(12), 1371–1375.

Tran, D., & Forsyth, D. (2007). Configuration estimates improve pedestrian finding. In *Advances in neural information processing systems*.

Viola, P., & Jones, M. (2002). Robust real-time object detection. *International Journal of Computer Vision*, *1*(2), 137–154.

Wang, Y., & Mori, G. (2008). Multiple tree models for occlusion and spatial constraints in human pose estimation. In *Proceedings of the European conference on computer vision*.