Upper Body Pose Estimation with Temporal Sequential Forests

James Charles[1]
j.charles@leeds.ac.uk
Tomas Pfister[2]
tp@robots.ox.ac.uk
Derek Magee[1]
d.r.magee@leeds.ac.uk
David Hogg[1]
d.c.hogg@leeds.ac.uk
Andrew Zisserman[2]
az@robots.ox.ac.uk

[1] School of Computing
University of Leeds
Leeds, UK

[2] Department of Engineering Science
University of Oxford
Oxford, UK

Abstract

Our objective is to efficiently and accurately estimate human upper body pose in gesture videos. To this end, we build on the recent successful applications of random forests (RF) classifiers and regressors, and develop a pose estimation model with the following novelties: (i) the joints are estimated sequentially, taking account of the human kinematic chain. This means that we don't have to make the simplifying assumption of most previous RF methods – that the joints are estimated independently; (ii) by combining both classifiers (as a mixture of experts) and regressors, we show that the learning problem is tractable and that more context can be taken into account; and (iii) dense optical flow is used to align multiple expert joint position proposals from nearby frames, and thereby improve the robustness of the estimates.

The resulting method is computationally efficient and can overcome a number of the errors (e.g. confusing left/right hands) made by RF pose estimators that infer their locations independently. We show that we improve over the state of the art on upper body pose estimation for two public datasets: the BBC TV Signing dataset and the ChaLearn Gesture Recognition dataset.

1 Introduction

The goal of this paper is to recover the 2D layout of human upper body pose over long video sequences. The focus here is on producing pose estimates for use in gesture analysis and recognition. As case studies, we show experiments in sign language and Italian gestures. In these domains we encounter gestures which can be performed continuously over up to an hour of video and with high degrees of variation in pose and person body shape. Furthermore, foreground appearance is highly varied with people wearing different clothing and standing in front of moving and cluttered background scenes.

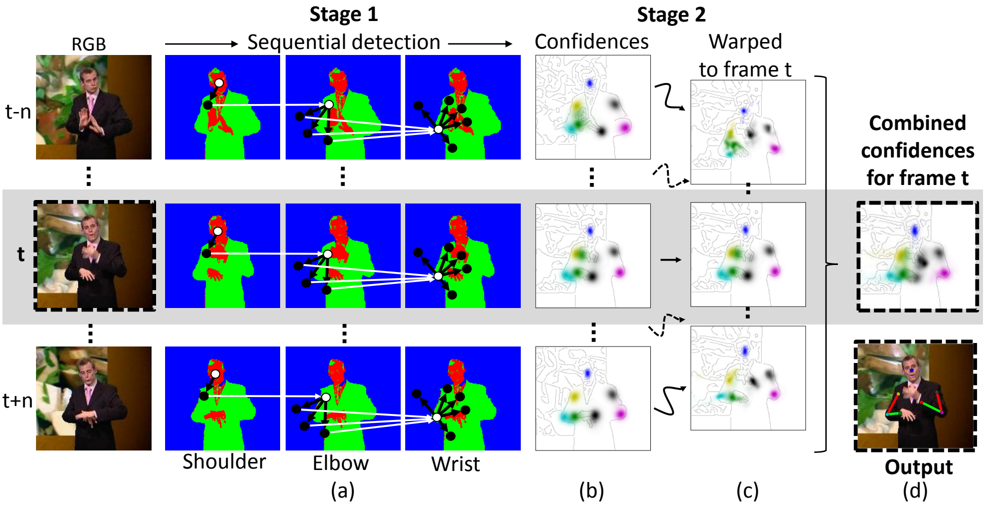


Figure 1: Process diagram of temporal sequential upper body pose estimation on RGB input frame t. Sequential detection of body joints (Stage 1) and detection reinforcement with optical flow (Stage 2) for a person's right arm is shown running from left to right. (a) Given the head position, joints are detected using a separate mixture of experts per joint, in order of shoulder, elbow then wrists. Knowledge of previous experts' output joint prediction (white arrows) is passed on when estimating the next joint in sequence. Each expert is responsible for a region of the image relative to the previous joint in sequence. Experts are positioned with fixed learnt offset vectors (positions shown as black dots with black arrows indicating offsets). (b) Joint detections produce a body joint confidence map (different colour per joint) shown superimposed on an edge image. (c) Confidences at frame t are reinforced by combining warped confidences (using optical flow) from n frames either side of t. (d) Body joint estimates are chosen as points of maximum confidence from the combined confidence map, with final pose output illustrated with a skeleton on RGB frame t.

We build upon recent work within the area of upper body pose estimation. The generative model by Buehler *et al.* [1] is capable of continuously tracking a person performing hours of sign language, but requires manual annotation of some frames, and is computationally expensive. This generative model was later used to train a much faster and more reliable pose estimator [2, 3, 13] using a sliding window random forest (RF). However, these methods only utilised context within a small sliding window per frame. They were unable to capture any global structure of the upper body and ignored dependencies between body parts. This resulted commonly in non-valid upper body poses, or confusions between left and right hand predictions. In our method we mitigate these problems by implicitly encoding an upper body kinematic chain using sequential forests. Furthermore, in the spirit of Zuffi *et al.* [20], we use dense optical flow as an additional observation to reinforce pose estimates at a given frame using neighbouring frames, thus leading to more constrained output, and improved accuracy.

RFs are a very powerful tool for pose estimation thanks to their comfortable scalability to very large amounts of training data and their computational efficiency. However, efficiency normally comes at the price of assuming independence for each output variable, and ignoring output structure [6, 15]. Past solutions to this have been of two kinds: post-processing methods, and implicit methods. Post-processing methods take the output of the RF and fit models to them, such as Markov or Conditional Random Fields [7, 11, 12], shape models to regularise local detections [4], or simply filtering the output by checking global consistency of local detections [18]. Usually post-processing methods are rather slow due to the additional overhead. In contrast, implicit methods build constraints between output variables into the detection method directly during training [8, 9, 10, 16] by passing output from a sequence of classifiers as input into another classifier. Recent developments in RF sequential classification are that of entangled decision trees [9], which were later extended by Kontschieder *et al.* [8] to use more context by including long-range features from geodesically filtering the class posteriors, and Dantone *et al.* [5], who fitted a pictorial structure model onto the output from two stacked RFs trained to approximate part dependencies.

The method presented in this paper (shown in Figure 1) addresses all of these issues by combining the benefits of both types of approaches (post-processing and implicit). First, unlike methods which aim to learn context, we take advantage of the kinematic constraints of the human body and explicitly build in context which we know is of importance, such as elbow location when detecting the wrist. A pose prior is implicitly built in by using a sequence of locally trained RF experts for each body parts (incorporating both classification and regression forests). This sequential detection method is capable of learning strong dependencies between connected body parts while keeping the search window small and the learning problem tractable. Second, our method removes the need for a sliding window part detector. This allows many more trees to be used, boosting accuracy while still maintaining efficiency. Third, our method's locally trained RFs deal with less of the feature space compared to its sliding window counterparts, which makes learning easier and, as we show, leads to improved accuracy. Fourth, temporal context is used to boost performance by incorporating output from multiple expert opinions from neighbouring frames. This is done using dense optical flow in a post-processing operation to warp output opinions to a single frame.

We evaluate our method on two datasets, the BBC TV signing sequences [2] and the ChaLearn gesture recognition dataset [14]. In both cases, our method is trained without any manual annotation of body poses. For the BBC TV dataset, ground truth poses are obtained using a generative upper body pose estimator [1]. For the ChaLearn dataset, we use the provided skeletal Kinect tracking output as training data. At test time we compare against manual ground truth and pitch our system against Kinect skeletal output.

2 Temporal Sequential Pose Estimation

We cast the task of upper body pose estimation as a detection problem. For each frame in the video, the configuration of the upper body parts is inferred by detecting body part joints: left/right shoulder, left/right elbow, left/right wrist and head centre.

Poses are estimated separately in two stages as shown in Figure 1: Stage 1, where body joints are detected sequentially; and Stage 2, where the detections are reinforced with temporal information from optical flow. We next give an overview of these stages.

Stage 1 – Sequential body joint detection. In this stage, body joints are detected sequentially in a single video frame. Each joint in the sequence depends on the location of the previous joint: the head is detected first, followed by shoulders, elbows, and wrists, separately for left and right arms. Stage 1 in Figure 1(a) illustrates this sequential detection. Beginning with an RGB frame, the frame is first encoded into a feature representation, shown in Figure 1(a) as an image with pixels categorised as either skin (red), torso (green) or background (blue). For each joint, a separate mixture of experts (random forests) votes for the next joint location (votes shown as white lines in figure). Each expert (shown as black dots in figure) is responsible for a particular region of the image dependent upon the location of the previous joint in the sequence (positioned according to fixed learnt offset vectors, shown as black arrows). The output from this process consists of a confidence map over pixels for each joint, illustrated in Figure 1(b) with each joint in a different colour, with higher-intensity colours indicating higher confidence.

Stage 2 – Detection reinforcement with optical flow. The joint detections in Stage 1 are done independently for each frame. However, in reality, strong dependencies exist between temporally close video frames. In this stage, confidences from Stage 1 produced at a frame t are reinforced with temporal context from nearby frames. Additional confidence maps are produced for neighbouring frames, and are then aligned with frame t by warping them backwards or forwards using tracks from dense optical flow. This is illustrated in Figure 1(c) for confidences produced within n frames either side of frame t. These confidences represent a strong set of expert opinions for frame t, from which joint positions can be more precisely estimated than when only using one confidence map. Finally, body joint locations are estimated at frame t by choosing positions of maximum confidence from a composite map produced by combining warped confidences (examples shown in Figure 1(d)).

Forest overview & comparison to prior art. Our sequential pose estimation method uses two types of random forests, one of each type, for each expert: (1) classification forests, which measure the likelihood an image region (determined by the learnt offset vector w.r.t. the previous joint) contains useful context for predicting the joint of interest, termed the region's 'utility score'; and (2) regression forests, which identify the joint's precise position within the regions, weighted by the utility scores from the classification forest. In contrast to the sliding window part detector RFs of [2, 5, 6, 15, 19], our expert joint detectors are constrained to local image regions assigned by the sequential RF. This considerably simplifies the learning task and yields a much more computationally efficient joint detector. In contrast to auto-context methods [16] and entangled forests [8, 9] which try to learn output dependencies, our offset vectors exploit prior knowledge about the human kinematic chain to explicitly encode much stronger dependencies between connected body parts.

3 Implementation Details

In this section we give further details about the two stages of our method.

3.1 Stage 1 – Sequential body joint detection

The sequential forest consists of five main components: (i) a global cost function, (ii) feature representation for input images, (iii) offset learning (between connected body parts), (iv) classification forests (for computing region utility scores), and (v) regression forests (voting for joint positions in the regions, weighed by their region utility scores). We describe each of these in detail.

Global cost function. The sequential detection model is encoded as two separate 'chains' of detections, one for each arm. Each detection chain maximises the output from a mixture of 'experts' (random forest regressors), each of which votes separately for the most likely position of a joint within a region. These experts' votes are weighted according to the region's 'utility score', provided by the random forest classifier. Given an input image I, upper body joint detection reduces to optimising the cost function:

$$l_i = \arg\max_{l_i} \sum_{k=1}^{K} w_{ik} p(l_i|W_{ik}), \qquad (1)$$

where $\{l_1, ..., l_4\}$ are 2D joint locations for head, shoulder, elbow and wrist for one of the arms; K is the number of experts (different for each body joint; we use 1 expert for shoulders, 3 for elbows and 5 for wrists); W_{ik} is the region for the k^{th} expert for joint i; w_{ik} are the region utility scores from the RF classifier, measuring how useful the context in image region W_{ik} is for predicting the position of the joint; and $p(l_i|W_{ik})$ is the likelihood of joint i being located at position l_i (obtained from the RF regressor). The locations of region centres are determined by previously detected joints and a learnt offset vector δ_{ik} (see below for how these are obtained) as $l_{i-1} + \delta_{ik}$.

Image representation. Both classification and regression RFs extract two types of features for each pixel. The first feature is a label which states whether the pixel belongs to a skin, torso or background region, obtained by maximising over a colour posterior image from our method from [2]. The second feature (inspired by Kontschieder *et al.* [8]) incorporates additional long-range context by measuring, for each pixel, the smallest Euclidean distance to a skin region (this is computed using a distance transform).

Learning expert offset vectors. The human kinematic chain sets clear physical constraints for where one joint (*e.g.* the elbow) can be located given another joint (*e.g.* the shoulder). These 'offsets' essentially represent the most usual positions of a joint relative to the previous joint in the sequence. We learn them from training data by clustering the relative offset of joint i from joint $(i-1)$ into K clusters using k-means (where K is defined above). The centroid of the clusters are used as the offsets δ_{ik}, and the variances are later used to add robustness to detections.

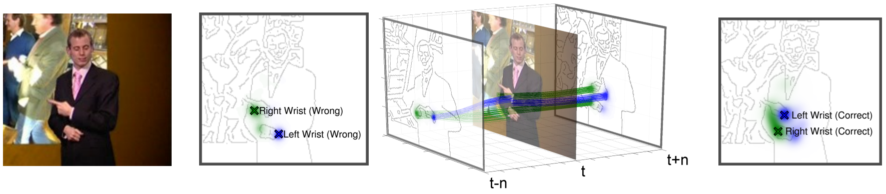
Classification forests. The offset vectors determine regions of the image that normally contain useful context for predicting the position of a given joint. However, for a given pose, not all pre-learnt regions are equally useful. We therefore employ a classification forest to obtain a 'utility score' for each of the regions. This utility score is used in the global cost function to weight the output from the expert body joint regressor (the regression forest). A separate classification forests, for each joint and expert, classifies an image region centred at W_{ik} as either *contained* or *not-contained* (indicating whether joint i is contained in the region or not). The average of all class probabilities across all trees is used to form the utility score w_{ik}. Parameters of test functions are learnt by minimising a measure of Gini impurity.

Regression forests. Given all pre-learnt regions and the position of the previous joint in the sequence, the task of the regression forests is to predict where in each region the joint is most likely to be. A separate regression forest, for each joint and expert, votes for pixel locations of joint i based on boolean tests performed on an image region centred at W_{ik}. All votes across all trees and all forests have equal weight. Parameters of test functions are learnt by recursively splitting a random selection of training regions (all of which contain the joint i) into two sets, based on minimising the sum of joint position variance within the two regions. The average joint position across regions falling at a leaf node is used as our voting vector. The aggregated voting vectors form the likelihood function $p(l_i|W_{ik})$.

In both forests: (1) robustness is added by classifying multiple regions shifted slightly by Gaussian noise, using variances from offset vector clustering, and taking the average over all regions; and (2) very efficient boolean test functions [2] are used.

3.2 Stage 2 – Detection reinforcement with optical flow.

The joint detections in Stage 1 are produced independently for each frame. However, in reality there exists very strong dependencies between poses in nearby frames. Therefore, to reinforce the joint confidences, we integrate confidences produced from the frame's neighbourhood. This is done in three steps: (1) the confidences from nearby frames are aligned to the current frame using dense optical flow; (2) these confidences are then integrated into



(a) Frame t (b) Confidence at t (c) Warping (d) Composite at t

Figure 2: Warping neighbouring confidence maps for improving pose estimates. (a) RGB input at frame t. (b) Confidence map at frame t for left (blue) and right (green) hands (crosses show incorrect modes of confidence). (c) Confidence maps from frames $(t-n)$ and $(t+n)$ warped to frame t using tracks from optical flow (green & blue lines). (d) Composite map with corrected modes.

a composite confidence map; and (3) the final upper body pose estimate for a frame is then simply the positions of maximum confidence from the composite map. Below we discuss the details of the first two steps.

Warping confidence maps with optical flow. For a given frame t, pixel-wise tracks are computed from neighbouring frames $(t-n)$ to $(t+n)$ to frame t using dense optical flow [17]. Tracks are used to warp confidence values within a neighbouring map to align them to frame t by shifting confidences along the tracks. Confidence values are averaged together at points where tracks merge. Example tracks and the warping of wrist confidence values are shown in Figure 2(c). These aligned confidences (Figure 1(c)) represent a strong set of expert opinions for frame t, from which joint positions can be more precisely estimated compared to using a confidence map from a single frame.

Composite confidence map. Aligned neighbouring confidence maps are integrated into a composite confidence map by taking the pixel-wise average (Figure 1(d) and Figure 2(d)). The composite map alleviates misdetections caused by failures in our feature representation. For example, Figure 3(a-d) shows examples where the image representation (b) loses relevant information w.r.t. the wrist location, with no hope of recovery using the confidences (c) from Stage 1. However, the composite confidence map (d) contains confidence in the correct regions even in these extreme circumstances. Figure 2(a-d) demonstrate the advantages of this method under the challenging scenario where the arms cross.

4 Datasets

Upper body pose estimation is evaluated on two video datasets: sign-interpreted BBC TV broadcasts, and an Italian gesture recognition dataset (see Figure 5(c) and Figure 4(e) for example frames). Both datasets contain large variations in poses, people, clothing and backgrounds, with the BBC TV dataset also containing moving backgrounds.

BBC TV sign-language broadcasts. This dataset consist of 20 TV broadcast videos overlaid with a person interpreting what is being spoken into British Sign Language (BSL) (see Figure 4(e)). The videos contain sign-interpreted content from a variety of TV programmes, each between 0.5h–1.5h in length. All frames of the videos have been automatically assigned joint locations (which we use as ground truth for training) using a slow but reliable tracker by Buehler *et al.* [1]. The full set of 20 videos are split into three disjoint sets: 10 videos for training, 5 for validation, 5 for testing. The test set videos contain different people and clothing from those in the training and validation sets. 200 annotated frames per video (1000 frames in total) of the test set have been manually annotated for evaluation purposes.

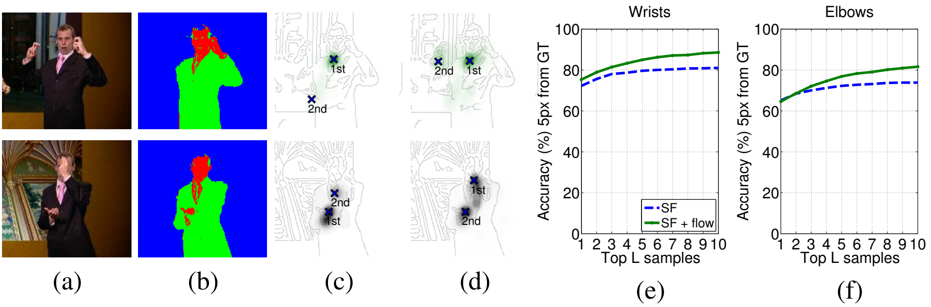


Figure 3: The improvement from using optical flow. (a) Example input images and (b) corresponding colour features, with the top row showing an error in segmentation (right arm missed), and bottom row showing a left hand in front of the face. (c) shows sequential forest (SF) confidences (superimposed on an edge image) for the left and right wrists for the two rows, with the first two modes of confidence (shown as crosses) both in wrong wrist locations. (d) demonstrates improved confidence output when combining SF output with optical flow (SF+flow): in the top row the 2nd best mode of confidence now correctly locates the right wrist, and in the bottom row the 1st mode of confidence locates the left wrist. Graphs in (e) and (f) show the accuracy for average wrist and elbow estimates (averaged over left and right body parts), where the accuracy measure reports the percentage of testing frames where at least one estimate out of L, chosen from top L modes of confidence, is within 5 pixels from ground truth.

ChaLearn gesture dataset. The ChaLearn 2013 Multi-modal gesture dataset [14] contains 23 hours of Kinect data of 27 persons performing 20 Italian gestures. The data includes RGB, depth, foreground segmentations and full body skeletons, and is split into training, validation and testing videos, each lasting 1-2min and containing 8-20 gestures. Example frames from the dataset are shown in Figure 5(c). The large variation in clothing across videos poses a challenging task for pose estimation methods.

Data pre-processing. For the BBC TV dataset, background subtractions and colour posterior images are produced for all video frames using the RGB frames and our method from [2]. From this we produce our feature representation. The same method is applied to the ChaLearn dataset, except we use the provided background subtracted images.

5 Evaluation

Four experiments are conducted using the BBC TV dataset. Experiments (1) and (2) evaluate the quality of the structured output of the sequential detection system, and experiments (3) and (4) evaluate the accuracy of pose estimates. Each experiment evaluates the two components our our system: (i) the sequential detection forests from Stage 1 (SF), and (ii) the detections reinforced using optical flow from Stage 2 (SF+flow). Experiment (5) evaluates our method on the ChaLearn gesture dataset using RGB frames, and compares against Kinect skeletal output. All experiments include comparisons to our state of the art upper body tracker from [2].

5.1 BBC TV experiments

Training. Each tree in both the classification and regression forests sample 500 diverse frames per training video (in total 5,000 frames for initial training, and 7,500 frames when using training and validation videos combined). Diverse frames are found by clustering frames with k-means (k = 200) according to pose, and sampling from clusters.

Testing. Pose estimates are evaluated against the 1,000 frames with manual ground truth.

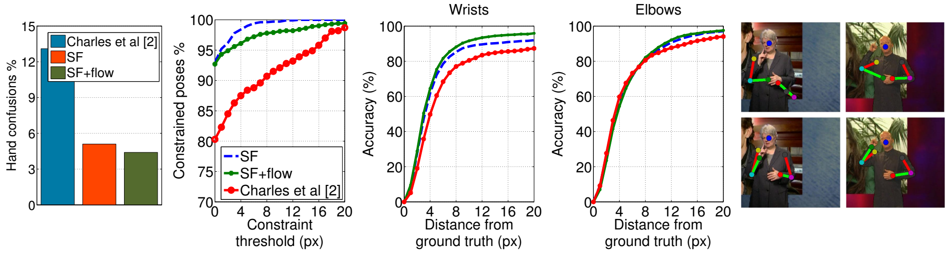
Parameter optimisation. Optimal parameters for the sequential detection forests are found on validation videos and set as follows. Classification forests: 8 trees per forest grown to a depth of 20 and using a square window size of 51. Regression forests: 8 trees per forest grown to a different depth per joint type, 20 for wrists, 10 for elbows and 20 for shoulders; a square window size of 31 pixels is used for all joint types. After fixing the parameters, the forests are retrained on a pooled set of all training and validation videos. For SF+flow we use a neighbourhood of $n = 15$ frames to compute composite confidence maps.

Baselines. We compare to our method from [2] using our new image features. Additionally, we test against another structured output method by implementing an auto-context [16] upper body detector. This auto-context method trains the forest of [2] using our image representation, and uses the output body joint confidences (together with the original image representation) as features to train a second forest. We also compare against the upper body pose estimator by Yang and Ramanan [19] (using code provided on authors website). For fairness we tune all baselines on the validation set and train with the same training material, except for estimator of Yang and Ramanan for which the INRIA dataset is additionally used to provide negative data.

Experiment 1: Constrained pose output. In this experiment we measure the proportion of output poses that are 'constrained' (essentially the proportion that is 'nearly correct'). A pose estimate is considered correctly constrained if the distance between connected joints (head to shoulder, shoulder to elbow and elbow to wrist) is less than the maximum ground truth projected limb length plus a threshold distance. Figure 4(b) shows the percentage of constrained pose outputs against a constraint threshold (in pixels). Notably, 20% of pose estimates from [2] have invalid limb lengths (at constraint threshold zero), whereas SF and SF+flow only have 7%, with SF completely eradicating unconstrained pose outputs at a constraint threshold of 13 pixels. A slight drop in constrained poses is seen when using SF+flow due to optical flow errors.

Experiment 2: Hand confusions. This experiment measures the proportion of frames in which the two hands of the humans are confused (*i.e.*, right hand is detected as left, and vice versa) – a very common error in previous methods. Hand confusions are detected by subjectively analysing output on the test set for frames where a wrist estimate is further than 5 pixels from ground truth. Manual labelling of hand confusion is used as automatic procedures are not reliable where the frames contain background segmentation errors or errors in the colour labelled image feature. As shown in Figure 4(a), SF reduces hand confusion errors by 61%, which is a considerable improvement over state of the art. SF+flow makes further improvements, reducing hand swap errors by 66%.

Experiment 3: Pose estimation accuracy. This experiment measures pose estimation accuracy as a function of distance from ground truth (a joint position is deemed correct if it lies within the set distance from ground truth). Figure 4(c) and (d) show estimated accuracy against allowed distance from ground truth for wrists and elbows. Improved accuracy is observed using SF over [2] for both wrists and elbows, with a particularly marked increase in accuracy for the wrists. Results at 5 pixels from ground truth for all applied methods are shown in the first two columns of Table 1(a). Using SF gives an improvement of 13% over [2]. SF+flow gives an additional boost of 15% for the wrists. Auto-context forests perform slightly better than [2], with worst performance from the pose estimator of Yang and Ramanan [19]. Due to the challenging nature of the dataset, we view the 15% improvement at wrist joints as a very significant improvement. It will be of great benefit to gesture recognition systems, where the location of hands is of utmost importance. SF without flow does not



(a) (b) (c) (d) (e)

Figure 4: Results on BBC TV dataset. (a) Bar chart showing percentage of hand confusions (lower is better) – SF+flow does best; (b) number of constrained (near-correct) poses across all values of the constraint threshold (higher is better); (c) & (d) accuracy of average wrist and elbows joints respectively (averaged of left and right arms) against allowed pixel distance from ground truth (higher is better) – SF+flow increases accuracy by 15% for wrists compared to [2] (at 5 pixels); (e) shows SF+flow (bottom row) correcting incorrect pose estimates from [2] (top row).

perform as well when (1) background subtraction fails (*e.g.* arms are cut off), or (2) hands are in front of the face or clothing contains skin-coloured regions. SF+flow alleviates these issues for the wrists, but less so for the elbows (due to frequent self-occlusions by the arms).

Experiment 4: Sampling joints. The pose estimator in this work produces a confidence map for possible joint positions in a frame (see Figure 3(d)). This experiment measures how accurate the joint predictions are if the correct sample out of the top N samples (instead of the maximum) from the confidence map is selected. Here we sample L locations from the top L modes of the confidence map as shown in Figure 3. A joint is deemed correctly estimated if any one of the L samples is within 5 pixels from ground truth. Figure 3(e) and (f) show plots of the accuracy for wrist and elbow joints vs different sample sizes. Using SF+flow introduces more modes of confidence and hence yields better results over using SF. The last two columns of Table 1(a) give results of SF and SF+flow when using 5 samples. A very significant improvement is observed over using just one sample (first two columns), with 85% wrist accuracy using SF+flow. This provides strong motivation for future research into producing a method for returning the best sample out of L.

5.2 ChaLearn experiments

Training & testing data. Each tree in both the sequential and classification forests sample 7,220 diverse frames from the training and validation videos combined. Diverse training frames are found by clustering poses and sampling uniformly from clusters. Testing data is formed by sampling 3,200 diverse frames from all testing videos. Kinect skeletal output is used as ground truth.

Parameter optimisation. We rescale the Chalearn videos to be of same size as those in the BBC TV dataset, and use the parameters found optimal in the BBC TV experiments.

Baselines. The upper body tracker from our previous work [2] is used as a baseline (trained using the same parameters as for the BBC TV dataset). We train with the same training frames used by our method.

Experiment 5: Pose estimation accuracy. Figure 5(a) and (b) show a comparison of the method of [2] and our sequential detection method (SF) without using optical flow. A significant improvement is seen in the wrists, with a stellar improvement for the elbows when using SF. Average accuracy at 5 pixels from ground truth is shown in Table 1(b). The

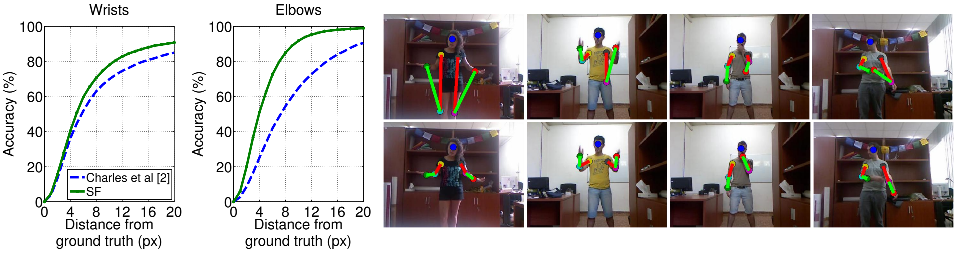
Method	Wrists	Elbows	Wrists - Top5	Elbows - Top5
Charles *et al.* [2]	60.5	67.6	-	-
Auto-context	62.1	68.7	-	-
Y & R [19]	42.9	41.7	-	-
SF	73.3	**70.1**	79.7	72.2
SF+flow	**75.3**	64.6	**85.0**	**76.9**

(a)

Method	Wrists	Elbows
Charles *et al.* [2]	44.7	33.6
SF	**50.8**	**62.6**

(b)

Table 1: Average accuracy for wrists and elbows at 5 pixels from ground truth. (a) Results for BBC TV (also showing accuracy of best estimates when sampling from top 5 modes). (b) Results for ChaLearn.



(a) (b) (c)

Figure 5: Joint detection results for Chalearn. (a) & (b) Average accuracy for wrists and elbows (against allowed distance from ground truth (Kinect provided skeleton) in pixels). Sequential detection forests (SF) show a significant improvement for wrist detections over [2], and nearly double the accuracy for the elbows at 5 pixels from ground truth. (c) Example pose estimates on Chalearn for [2] (top row) and SF (bottom row).

method of [2] does not generalise well to persons wearing trousers and sleeves of different length (due to confusions caused by legs and arm skin regions). The constrained output of SF helps overcome these problems. Finally, Figure 5(c) shows example pose estimate from [2] and SF in top and bottom rows respectively.

Computation time. Using a 2.2GHz Intel Quad Core I7 CPU, computation time on a 320x202 pixel image is: 0.4s for the sequential forest (272 trees in total); 0.14s for background segmentation and computing the image representation and 1.2s for computing optical flow. The sequential forest is implemented in Matlab with each tree only taking 1.5ms to evaluate the whole image due to the reduced search space. A classification tree takes 2.5 hours to train and a regression tree takes 4.5 hours.

6 Conclusion

A method has been proposed for tracking the upper body pose of people performing gestures in video captured using a monocular colour camera. We have addressed the problems of obtaining output dependences between body parts using a random forest pose estimator, and also exploit temporal context in a novel way using dense optical flow to spatially align sensory information through time. Significant increase in accuracy for wrist and elbow joints is observed on two separate challenging datasets compared with the state of the art – these joints are particularly important in gesture recognition. In future work we will investigate methods of returning better joint estimates by sampling from output confidences.

Acknowledgements: We are grateful to Patrick Buehler for his generous help. Financial support was provided by the EPSRC grants EP/I012001/1 and EP/I01229X/1.

References

[1] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Upper body detection and tracking in extended signing sequences. *IJCV*, 2011.

[2] J. Charles, T. Pfister, M. Everingham, and A. Zisserman. Automatic and efficient human pose estimation for sign language videos. *IJCV*, 2013.

[3] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman. Domain adaptation for upper body pose tracking in signed TV broadcasts. In *Proc. BMVC*, 2013.

[4] T.F. Cootes, M.C. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *Proc. ECCV*, 2012.

[5] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *Proc. CVPR*, 2013.

[6] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *Proc. ICCV*, 2011.

[7] J. Jancsary, S. Nowozin, T. Sharp, and C. Rother. Regression tree fields - an efficient, non-parametric approach to image labeling problems. In *Proc. CVPR*, 2012.

[8] P. Kontschieder, P. Kohli, J. Shotton, and A. Criminisi. Geof: Geodesic forests for learning coupled predictors. In *Proc. CVPR*, 2013.

[9] A. Montillo, J. Shotton, J. Winn, J. E. Iglesias, D. Metaxas, and A. Criminisi. Entangled decision forests and their application for semantic segmentation of ct images. *IPMI*, 2011.

[10] D. Munoz, J. A. Bagnell, and M. Hebert. Stacked hierarchical labeling. In *Proc. ECCV*, 2010.

[11] Payet N. and Todorovic S. $(\mathrm{RF})^2$ - random forest random field. In *NIPS*, 2010.

[12] S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kohli. Decision tree fields. In *Proc. ICCV*, 2011.

[13] T. Pfister, J. Charles, M. Everingham, and A. Zisserman. Automatic and efficient long term arm and hand tracking for continuous sign language TV broadcasts. In *Proc. BMVC*, 2012.

[14] Escalera S., J. Gonzalez, X. Baro, M. Reyes, O. Lopes, I. Guyon, V. Athistos, and H.J. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *ICMI*, 2013.

[15] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proc. CVPR*, 2011.

[16] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE PAMI*, 2010.

[17] P. Weinzaepfel, Z. Revaud, J.and Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *Proc. ICCV*, 2013.

[18] H. Yang and I. Patras. Sieving regression forest votes for facial feature detection in the wild. In *Proc. ICCV*, 2013.

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

[20] S. Zuffi, J. Romero, C Schmid, and M.J. Black. Estimating human pose with flowing puppets. In *Proc. ICCV*, 2013.