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

The goal of this work is to recover the 2D layout of human upper body pose over long video sequences. The focus is on producing reliable and accurate pose estimates for use in gesture analysis and recognition.

We build on the recent successful applications of random forests (RF) classifiers and regressors [1], 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 processing steps are divided into two stages.

Stage 1 – Sequential body joint detection

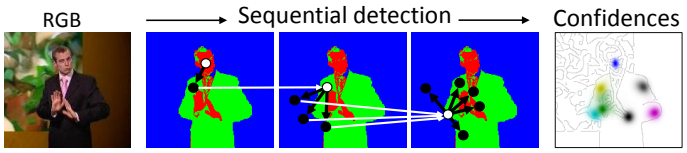
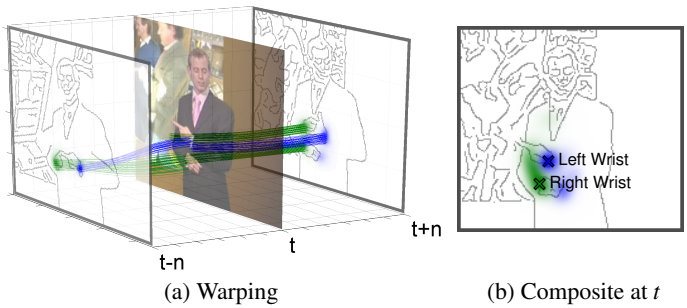


Figure 1: **Stage 1 – sequential upper body pose estimation.** (a) RGB input. (b) Sequential detection with random forest experts: the head is detected first, then shoulders, elbows and finally wrists. (c) Confidence map of body joints, with different colour for each joint (higher colour intensity indicates stronger confidence).

In Stage 1, 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. Figure 1(a-c) illustrates this sequential detection. Beginning with an RGB frame (a), the frame is first encoded into a feature representation, shown in Figure 1(b) as an image with pixels categorised as either skin (red), torso (green) or background (blue). For each joint, a separate mixture of experts (random forest) 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 which depends 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 is a confidence map over pixels for each joint.

Stage 2 – Detection reinforcement with optical flow



(a) Warping (b) Composite at t

Figure 2: **Stage 2 – warping neighbouring confidence maps to improve wrist joint detections.** (a) Confidence maps from frames $(t-n)$ and $(t+n)$ warped to frame t using tracks from optical flow (green & blue lines). (b) Composite map with crosses indicating modes of confidence.

In Stage 2, 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 aligned with frame t by warping them backwards or forwards using tracks from dense optical flow. This is illustrated in Figure 2(a) for wrist confidences produced at frame $(t-n)$ and $(t+n)$. Finally, body joint locations are estimated at frame t by selecting positions of maximum confidence from a composite map produced by combining warped confidences (see Figure 2(b)).

Figure 3: Pose estimates from our method on two different gesture datasets. Top: BBC TV dataset. Bottom: Chalearn gesture dataset.

Results

Our method takes advantage of the kinematic constraints of the human body and explicitly builds in spatial context which we know is of importance, such as elbow location when detecting the wrist. The locally trained RFs deal with less of the feature space compared to its sliding window counterparts, which makes learning easier and leads to improved accuracy over the state-of-the-art [1, 2].

Accuracy of the sequential forest at Stage 1 (SF) is shown to improve further when incorporating output from multiple expert opinions from neighbouring frames in Stage 2 (SF+flow) (see Figure 4). Example pose estimates on two different datasets are shown in Figure 3.

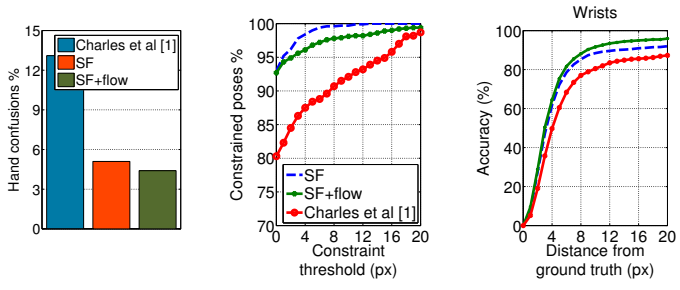


Figure 4: (a) SF+flow significantly reduces hand confusions. (b) SF and SF+flow achieve significantly better constrained pose estimates than state-of-the-art [1]. (c) Improvement in average wrist accuracy.

References

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

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