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

Tomas Pfister[1]
tp@robots.ox.ac.uk
James Charles[2]
j.charles@leeds.ac.uk
Mark Everingham[2]
m.everingham@leeds.ac.uk
Andrew Zisserman[1]
az@robots.ox.ac.uk

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

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

We present a fully automatic arm and hand tracker that detects joint positions over continuous sign language video sequences of more than an hour in length.

The standard approach of Buehler *et al*. [1] for tracking arms and hands requires manual labelling of 64 frames per video, which is around three hours of manual user input per one hour of TV footage. In addition, the tracker (by detection) is based on expensive computational models and requires hundreds of seconds computation time per frame. These two factors have hindered the large scale application of this method. In this paper we describe a method for tracking joint positions (of arms and hands) without any manual annotation and, after automatic initialisation, the system runs in real-time.

Our contributions are (i) a co-segmentation algorithm that automatically separates the signer from any signed TV broadcast using a generative layered model; (ii) a method of predicting joint positions given only the segmentation and a colour model using a random forest regressor; and (iii) demonstrating that the random forest can be trained from an existing semi-automatic, but computationally expensive, tracker. Figure 1 illustrates the processing steps.

Random forests for pose estimation. In recent years there has been increasing interest in random forest/fern-based methods. In particular we are interested in the work on human pose estimation, where random forests have most recently been used to infer full body pose [2]. However, the success of these pose methods depends upon the use of depth imagery which is colour and texture invariant, while also making background subtraction much easier. Here we propose an upper body pose estimation method that exploits the efficiency and accuracy of random forests without the need for depth images, and instead use raw RGB images with only a partially known background (as described below). See Figures 4 and 5 for illustrations of this method and a comparison against ground truth.

Co-segmentation for signer extraction. Co-segmentation methods consider sets of images where the appearance of foreground and/or background share some similarities, and exploit these similarities to obtain accurate foreground-background segmentations. In our case we exploit the fact that sign language broadcasts consist of a layered model of the foreground and two separate backgrounds, one that is static throughout each video and another that changes with each frame. The signer stands partially against the static background and partially against the changing background (which they are describing).

To this end we propose a co-segmentation algorithm that automatically separates signers from any signed TV broadcast by building a generative layered model as shown in Figures 2 and 3. We use this layered model of the signer in conjunction with a foreground colour model to provide a suitable input representation for the random forest regressor, superior to using the raw input image itself, and not requiring depth data.

The method is applied to signing footage with changing background, challenging imaging conditions, and for different signers. We achieve superior joint localisation results to those obtained using the method of Buehler *et al*. [1].

[1] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Upper body detection and tracking in extended signing sequences. *IJCV*, 95(2):180–197, 2011.

[2] 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.

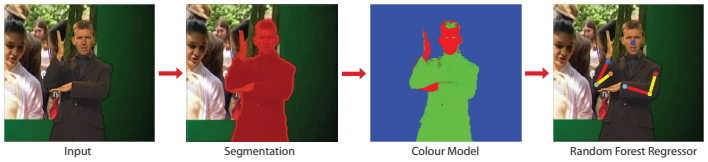


Figure 1: Arm and hand joint positions are predicted by first segmenting the signer using a layered foreground/background model, and then feeding the segmentation together with a colour model into a random forest.

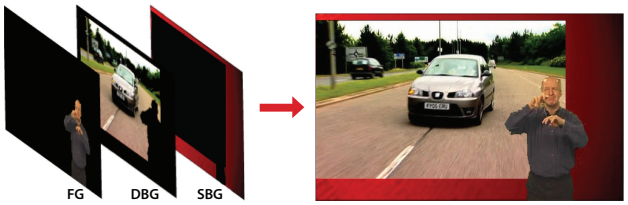
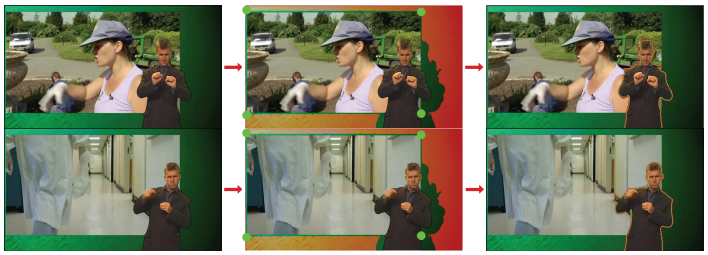
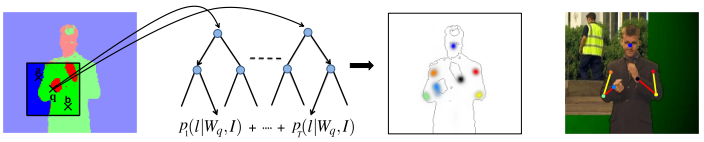


Figure 2: Generative layered model of each frame. The co-segmentation algorithm separates the signer from any signed TV broadcast by building a layered model consisting of a foreground, dynamic and static background.



(a) (b) (c)

Figure 3: Co-segmentation. (a) the original frames; (b) the changing background (rectangle spanned by the green dots) and the permanently fixed background (in red); (c) the final segmentation.



(a) Color model (b) Random forest (c) PDF of joints (d) Estimated joints

Figure 4: Estimating joint positions. (a) input colour model image; (b) random forest classifies each pixel using a sliding window; (c) probability density function of each joint location, shown in different colours per joint (more intense colour implies higher probability); (d) joint estimates, shown as small circles linked by a skeleton.

(a) (b) (c)

Figure 5: Joint estimation results. (a) shows a colour posterior from which we obtain probability densities of joint locations in (b) (black crosses mark maximum probability); (c) compares estimated joints (filled circles) with ground truth (open circles).