Upper Body Pose Estimation with Temporal Sequential Forests

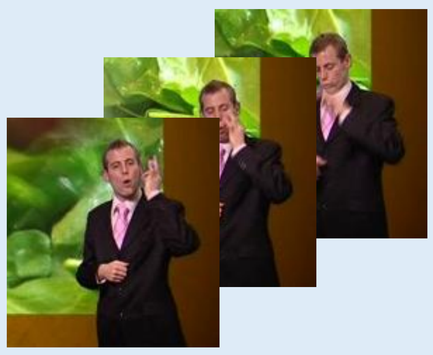
UNIVERSITY OF OXFORD [1]

J. Charles[2], T. Pfister[1], D. Magee[2], D. Hogg[2], A. Zisserman[1]

UNIVERSITY OF LEEDS [2]

Goal

Our goal is to obtain reliable **2D upper body pose estimation** over long video sequences in real-time for gesture recognition.



BBC TV Sign Language

Italian Gesture

Contributions

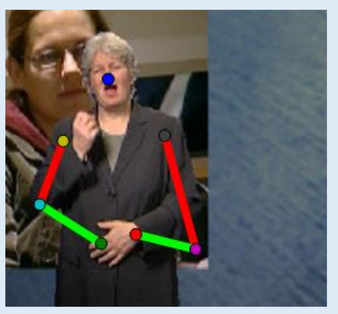
1. A random forest framework for learning **body joint dependencies**
2. Automatic identification of **meaningful image context**
3. Tractability of using a **mixture of random forest experts**
4. Temporal reinforcement using **dense optical flow**

Motivation

Address problems with **independent joint detection**

Unconstrained pose output

Ambiguities between hand positions

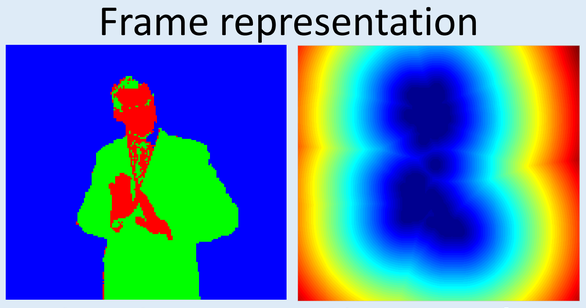


Solution: **Temporal Sequential Forests**

1 - Frame representation

RGB input

Frame representation

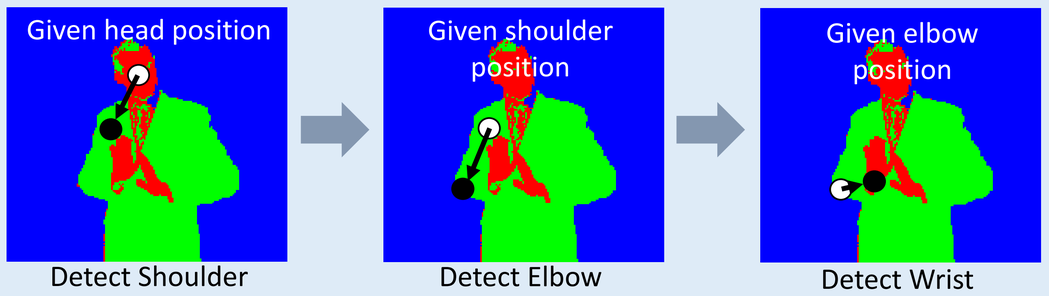


Skin/Torso/Background labelled image

Distance transform on skin regions

2 – Sequential forest detection (SF)

Each body joint is **conditionally detected** using a **mixture of experts**

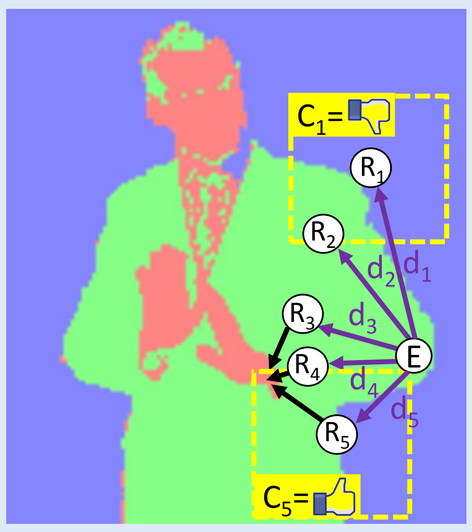


Detect Shoulder | Detect Elbow | Detect Wrist

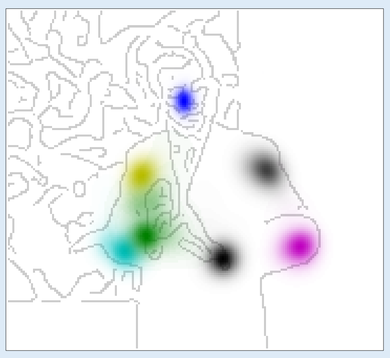
The joints are detected in sequence for each arm separately

3 – Mixture of random forest experts

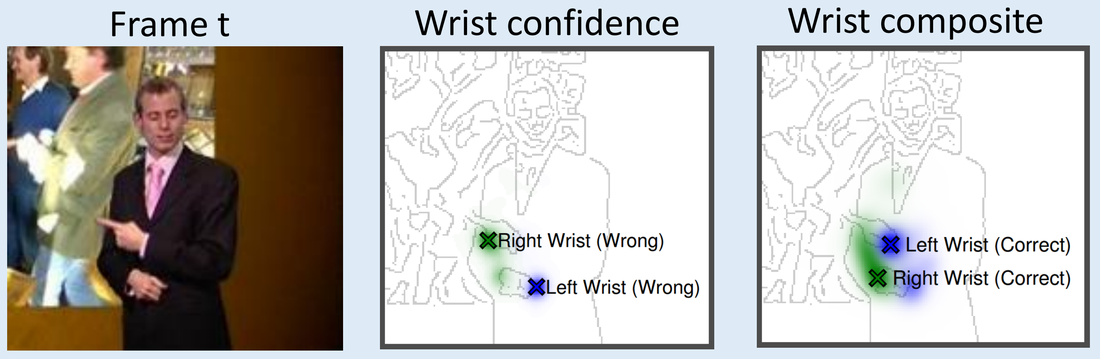
Example **wrist detection** given elbow position



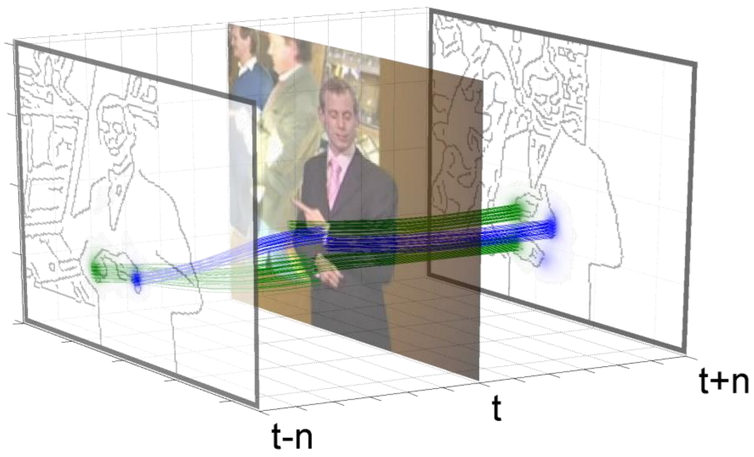
1. Given location of **elbow** (E)
2. Position K **regression forest experts** (R_i) according to learnt offsets d_i
3. Image context around each expert is scored according to its usefulness by a **classification forest** C_i
4. Experts vote for elbow position using these scores as a voting weight.
5. Votes are accumulated for each body joint to produce confidence maps. Different colour per joint, higher intensity colour means higher confidence



4 – Reinforcement with flow (SF+flow)



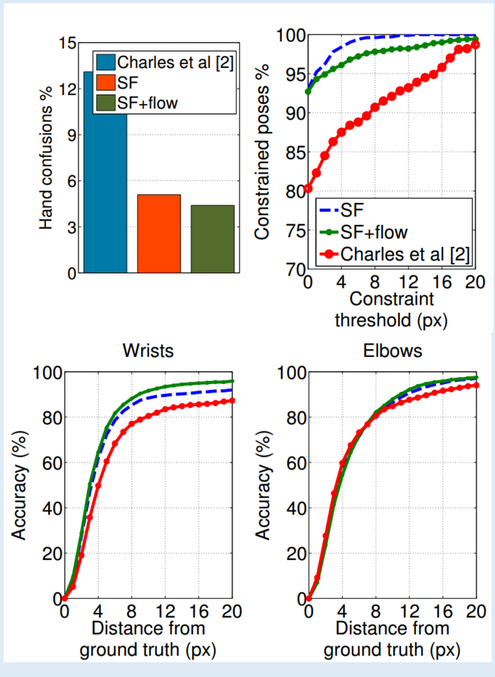
Confidence map at frame t is **corrected by a composite map**



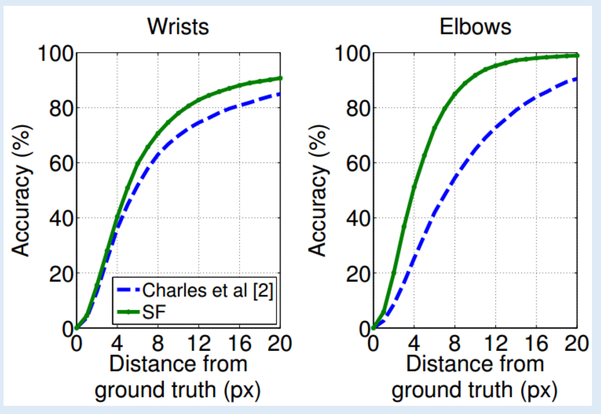
Composite map formed by **warping confidence** values in neighbouring frames along tracks produced from **dense optical flow** and summing them pixel-wise at frame t

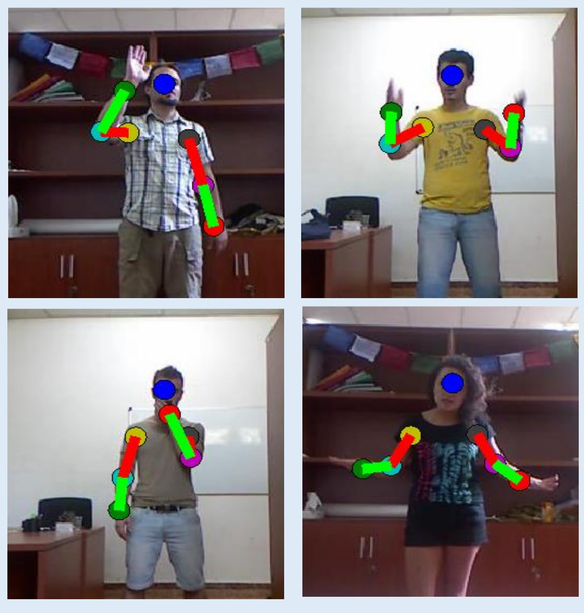
Results

BBC TV dataset [2]



ChaLearn 2013 Multi-modal gesture dataset [1]





References

[1] 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

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