“Hello! My Name is … Buffy” – Automatic naming of characters in TV Video [Everingham et al, BMVC 2006]

This paper aims at automatic labeling appearances of characters in TV or film videos.

The major steps of the method include:

1. **Subtitle and script processing**
   Subtitles record what is said but not by whom. Script records who says what, but not when. Subtitles and scripts are aligned using dynamic programming string alignment algorithm. Subtitles are extracted by OCR based “SubRip” program and spell corrected. Scripts are obtained from fan site and are preprocessed to extract text.

2. **Video Processing**

   a. **Face Detection and tracking**
      - Frontal face detection algorithm to avoid false positives.
      - Shot changes are detected based on color histograms.
      - KLT tracker is run on each track to obtain set of point tracks starting at some and continuing to some later frame.
      - Number of point tracks passing through two faces is used to decide if faces match.
      - Face detector output gives approximate locationand scale of face.

   b. **Facial Feature Localization**
      - Nine facial features are located in the face region using generative model of feature positions combined with a discriminative feature appearance
      - PDF of locations : mixture of Gaussian trees
      - Appearance of facial features: Haar like features + Adaboost algorithm

   c. **Representing Face Appearance**
      - Compute SIFT features around located facial features
      - Prior to that face region is geometrically normalized – reduces scale uncertainty, effect of pose variation

   d. **Representing Clothing appearance.**
      - YCbCr color space descriptors compared using chi squared distance

- This is a secondary factor, to improve identification – cannot be used by itself.
- Clothes within facial bounding box are used

e. Speaker Detection

- The person speaking might not be in the video frame e.g: reaction shot
- Identify lip movement : difference in the mouth region between current frame and next.

3. Classification by Exemplar Sets

- Tracks with single unambiguous identity are chosen as exemplars
- Quasi-likelihood that face corresponds to a particular name is defined and minimum distance between the descriptors in F and the exemplar tracks is defined and posterior probability is computed by Bayes rule.

Results: Evaluation against baseline.

Talking Heads: Detecting Humans and Recognizing Their Interactions [Minh Hoai, Andrew Zisserman]

Certain configurations of people appear more often in TV shows. This work aims to make use of this fact to provide scene context.

- Upper Body Configuration Detector (UBC) takes an image frame and outputs a configuration of upper body, specifying their location and scale.
- Use ensemble of UBC models from exemplar configuration
- Build configuration clusters (CC): A set of exemplar configurations for 2 or more UBs is obtained by two level hierarchical clustering:
  - Configuration vector of the UBS w.r.t UB union.
  - Configuration vector of the UB union w.r.t reference frame
- For a given video goal is to detect all UBs and recognize their configuration.
- Set up energy minimization problem: we want a set of UBs
  - High detection scores (unary potentials)
  - Low deformation cost w.r.t CC (prior)
- Dependency between UBs and union and that between union and image frame is a tree structure => dynamic programming and generalized distance transforms for searching all possible configurations in an image.
- Energy minimization is solved using max margin learning (convex quadratic). (Cplex)
- To detect background people, singleton detector is employed.
- Used on following tasks:
  - upper body detection and counting (outperforms DPM)
  - human interaction recognition (outperforms DTD) (human focused descriptor computed by averaging all UB track descriptors

+ HOG based scene descriptors : average of HOG descriptors computed on 3 key frames in shot)

- Datasets: TV Human Interaction dataset (TVHI) and 150 episodes dataset