High Five: Recognising human interactions in TV shows

Alonso Patron-Perez
alonso@robots.ox.ac.uk
Marcin Marszalek
marcin@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk
Ian Reid
ian@robots.ox.ac.uk

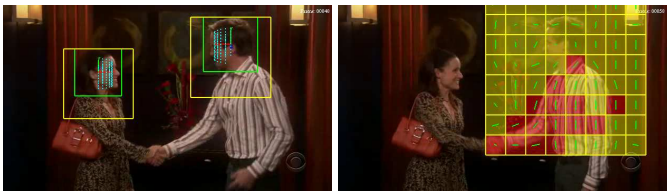
Department of Engineering Science
University of Oxford
Oxford, UK

The aim of this paper is the recognition of *interactions* between two people in videos in the context of video retrieval. In contrast to previous work in this area, we test our method with a more realistic dataset that we have compiled from TV shows[1]. This dataset contains examples of four interactions: hand shakes, high fives, hugs and kisses, as well as negative examples (Figure 1).

Figure 1: Dataset snapshots. Note the variation in the actors, scale and camera views.

An upper body detector [3] is first used to find people in every frame of the video. The detections are then clustered to form tracks. A *track* is defined as a set of upper body bounding boxes, in consecutive frames, corresponding to the same person. This first step reduces the search space for interactions to a linear search along each track [4].

A person's local context is described by superimposing an 8×8 grid around an upper body detection and calculating histograms of gradients and optical flow in each of its cells. These histograms together with the head orientation are used to create a descriptor. The head orientation is discretised into one of five orientations [1]. The local context aims to capture cues such as hand and arm movement (Figure 2b), while the head orientation is used to capture weak correlations between the local context and the camera view. Using these descriptors, a one-vs-the-rest linear SVM classifier is learnt for each interaction.



(a) (b)

Figure 2: (a) Upper body detections and estimated discrete head orientation. (b) Grid showing dominant cell gradient (green lines) and significant motion (red cells) for a hand shake.

To improve the classification obtained with the person-centred descriptor, we assume that people face each other while interacting (Figure 2a). The goal is to simultaneously estimate the best joint classification for a set of detections in a video frame rather than classifying each detection independently. The matching cost of using a joint labeling for a set of frame detections takes into account the SVM interaction scores of each independent detection and the relative location of people in the frame discretised into one of six spatial relations shown in Figure 3. Learning is done using structured SVM [2, 5], and the label that maximises the matching cost is found by exhaustive search.

We tested various modifications of the person-centred descriptor and evaluated the structured learning method in a video retrieval task performed with our dataset obtaining promising results (Figure 4). Our conclusion is that using head orientation and structured learning can improve the classification of interactions in a significant way, but a reliable head pose classifier and upper body detectors are necessary to maximise this improvement.

[1] http://www.robots.ox.ac.uk/∼vgg/data/tv_human_interactions

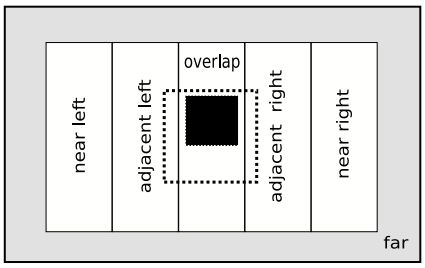


Figure 3: Spatial relations used in our SL method. The black square at the centre represents the head location inside an upper body detection.

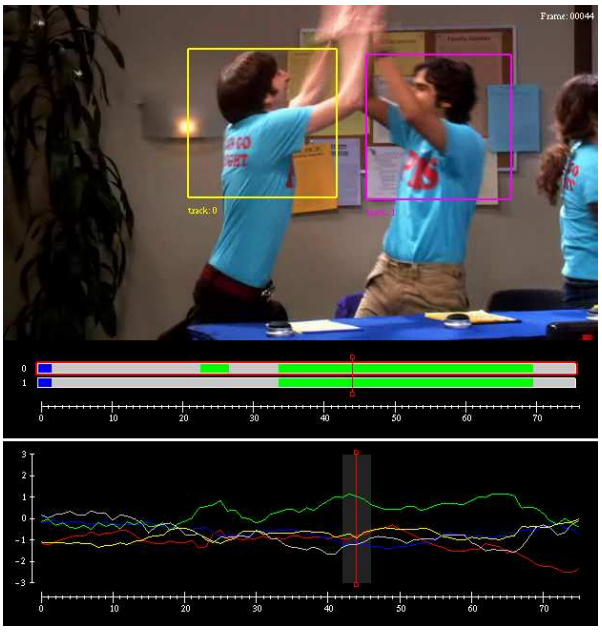


Figure 4: Snapshot of a clip with two tracks. Frame classification and SVM interaction scores are shown (where green represents **high five** and gray no interaction).

[1] B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *BMVC*, 2009.

[2] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009.

[3] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Pose Search: retrieving people using their pose. In *CVPR*, 2009.

[4] A. Kläser, M. Marszalek, C. Schmid, and A. Zisserman. Human Focused Action Localization in Video. In *SGA*, 2010.

[5] I. Tsochantaridis, T. Hofman, T. Joachims, and Y. Altun. Support Vector Machine Learning for Interdependent and Structured Output Spaces. In *ICML*, 2004.