"Here's looking at you, kid." Detecting people looking at each other in videos.

Manuel J. Marín-Jiménez
mjmarin@uco.es
Andrew Zisserman
az@robots.ox.ac.uk
Vittorio Ferrari
ferrari@vision.ee.ethz.ch

Dpt. of Computer Science and Numerical Analysis
University of Cordoba, Cordoba, Spain
Dpt. of Engineering Science
University of Oxford, Oxford, UK
Dpt. of Information Technology and Electrical Engineering
ETHZ, Zurich, Switzerland

If you read any book on film editing or listen to a director's commentary on a DVD, then what emerges again and again is the importance of eyelines. Standard cinematography practice is to first establish which characters are looking at each other using a medium or wide shot, and then edit subsequent close-up shots so that the eyelines match the point of view of the characters. This is the basis of the well known 180^o rule in editing.

Figure 1: **Are they looking at each other?** Answering this question enables richer video analysis, and retrieval based on where actors interact. From left to right: *Friends, Casablanca, Fawlty Towers*.

The objective of this paper is to determine whether eyelines match between characters within a shot – and hence understand which of the characters are interacting. The importance of the eyeline is illustrated in the examples of fig. 1 – one giving rise to arguably the most famous quote from *Casablanca*, and another being the essence of the humour at that point in an episode of *Fawlty Towers*. Our target application is this type of edited TV video and films. It is very challenging material as there is a wide range of human actors, camera viewpoints and ever present background clutter. Determining whether characters are interacting using their eyelines is another step towards a fuller video understanding. Our approach can be summarized in three stages:

1. Human head detection and tracking. We start by detecting human heads in video shots and grouping them over time into tracks, each corresponding to a different person. The head detection process is split into two subprocesses. First, an upper-body detector is applied to each frame independently and these detections are then grouped over time into tracks. Second, heads are detected within each upper-body detection and, again, these detections are grouped into tracks.

2. Head pose regression. Next, we estimate the *pitch* (around X axis) and *yaw* (around Y axis) angles for each head detection. For this, we follow a 2D approach and train a Gaussian Process regressor [2] using publicly available datasets to estimate the head pitch and yaw directly from the image patch within a detection window. The input to the Gaussian Process regressor is a HOG descriptor of the head region.

3. Classifying pairs of head tracks as looking at each other. For the last stage, we investigate three methods to determine if two people (head tracks) are looking at each other (*LAEO*). Two people are LAEO if there is eye contact between them. We start with a simple 2D analysis, based on the intersection of gaze areas in 2D defined by the sign of the estimated yaw angles. In a more sophisticated alternative, we use both the continuous yaw and pitch angles as well as the relative position of the heads. Finally, we propose a '2.5D' analysis, where we use the scale of the detected head to estimate the depth positioning of the actors, and combine it with the full head pose estimate to derive their gaze volumes in 3D (see fig. 2).

Results. We apply these methods to the TV Human Interactions Dataset (TVHID) [1]. This is very challenging video material with far greater variety in actors, shot editing, viewpoint, locations, lighting and clutter than the typical surveillance videos used previously for classifying interactions where there is a fixed camera and scene. TVHID contains a total of 300 video clips grouped in five classes: *hand-shake*, *high-five*, *hug*, *kiss* and *negative*. The dataset is divided equally into training and test sets. For our task, we have additionally annotated all the videos by assigning one of the following labels to each shot *label 0*: no pairs of people are LAEO; *label 1*: one or more pairs of people are LAEO in a clearly visible manner; *label 2*: a pair of people are LAEO, but at least one of them has occluded eyes (e.g. due to viewpoint or hair); and, *label 3*: a pair of people are facing each other, but at least one of them has closed eyes (e.g. during kissing). There is a total of 443 video shots, where 112 have label 0, 197 label 1, 131 label 2 and 3 label 3. Therefore, the dataset contains 112 negative (label 0) and 331 positive samples (labels 1, 2 and 3).

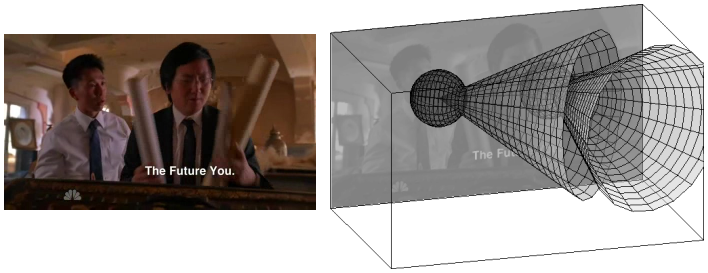

Figure 2: **Geometric constraints in 3D. (left)** Original video frame. **(right)** 3D representation of a scene with two people. We show the person's head (spheres) and their gaze volumes (cones).

We evaluate here the performance of the proposed LAEO classifiers on the following task: *is there any pair of people LAEO at any time in this video shot?* Each method is used to score every shot, and then we measure its performance as the average precision (AP) over the test set. We run experiments on two trials, where one partition of the TVHID is used for training and the other for testing, and then report mean AP (mAP) over the two trials.

Figure 3: **Top LAEO shots.** Note the variety of clothing, backgrounds, and illuminations that are handled by the method.

If all test shots are scored with uniformly distributed random values, the mAP over 10 trials is 0.75. We define a baseline method based on the directional information provided by our head detector. This achieves a mAP of 0.816 (i.e. better than chance). However, our best method (i.e. geometric constraints in 3D) achieves a mAP of 0.862. Figure 3 shows six examples of top ranked shots, retrieved by our best method, highlighting the variety of situations where it succeeds. While we report quantitative performance at the shot level, our method localizes the interacting people both spatially (i.e. the pair of heads with the highest LAEO score) and temporally (thanks to a temporal sliding window mechanism). In conclusion, the recognition of LAEO pairs introduces a new form of high-level reasoning to the broader area of video understanding.

We publicly release the LAEO annotations at the following URL: `http://www.robots.ox.ac.uk/~vgg/data/`.

[1] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman. High Five: Recognising human interactions in TV shows. In *Proc. BMVC.*, 2010.

[2] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.