Davies, S. J. C., Agrafiotis, D., Canagarajah, C. N., & Bull, D. R. (2008). A gaze prediction technique for open signed video content using a track before detect algorithm. In 15th IEEE International Conference on Image Processing, 2008 (ICIP 2008). (pp. 705 - 708). Institute of Electrical and Electronics Engineers (IEEE). 10.1109/ICIP.2008.4711852

A GAZE PREDICTION TECHNIQUE FOR OPEN SIGNED VIDEO CONTENT USING A TRACK BEFORE DETECT ALGORITHM

S.J.C. Davies, D. Agrafiotis, C.N. Canagarajah and D.R. Bull

Department of Electrical and Electronic Engineering
University of Bristol, Bristol, BS8 1UB, UK
e-mail: sam.davies@bris.ac.uk

ABSTRACT

This paper proposes a gaze prediction model for open signed video content. A face detection algorithm is used to locate faces across each frame in both profile and frontal orientations. A grid-based likelihood ratio track before detect routine is used to predict the orientation of the signer's head, which allows the gaze location to be localised to either the signer or the inset. The face detections are then used to narrow down the gaze prediction further. The gaze predictor is able to predict the results of an eye tracking study with up to 95% accuracy, and an average accuracy of over 80%.

Index Terms— Video contexts, video coding, gaze tracking

1. INTRODUCTION

Open sign language is the term used to describe video with an in vision signer and is used as an alternative to subtitles, which are often inadequate at conveying concecpts, such as emotion. Figure 1 shows a sample frame from open signed video footage.

Fig. 1: A sample frame from one of the '*holby*' open signed video sequences

Open sign language is an excellent candidate for perceptual video coding due to the well-defined nature of the gaze patterns, as shown in [1], [2], [3] and [4]. Previous work has been done on analysing the gaze patterns of sign language in various forms, including videoconferencing [4] and open sign video [2], and both went on to code video with perceptually varying quality. It was shown that within the sign language context it is possible to make coding gains without loss of intelligibility or perceived quality.

In order to improve on these techniques we look at improving the gaze prediction routine, specifically for open sign lanugage. Previous work into gaze prediction has often involved the concept of saliency [5], where low-level features of the video (such as colour contrast, intensity, orientation etc.) are combined into a saliency map, and this is then used to plot predicted gaze patterns. Due to the bottom-up nature of this scheme it has been shown to be inadequate for many specified contexts, sign language included [2].

An eye tracking study was carried out (see Section 2) which shows that the orientation of the signer's head is an excellent cue as to whether or not the viewers would be looking at the signer or the inset video. Therefore we first propose a grid-based likelihood ratio tracking system to predict the orientation of a face (i.e. frontal or profile), before utilising this tracker to construct a gaze prediction routine for open signed video. The gaze predictor outputs the most likely locations of an eye fixation for each frame. These locations can be used for various applications including to modulate a variable quality coder, or add region-based error protection.

This paper is organised as follows: Section 2 details an eye tracking study carried out with open sign language material, including, in Section 2.1, some observations of the collected data. Section 3 introduces the proposed gaze prediction technique, with the face detector being detailed in Section 3.1, the track before detect routine in Section 3.2, and the gaze predictor in Section 3.3. Section 4 demonstrates results from both the tracker and the gaze predictor, before the paper is concluded in Section 5, including a discussion on further work.

2. EYE TRACKING STUDY

Five participants, each of whom is fluent in British Sign Language (BSL), were shown a random selection of open signed clips at standard definition (720×576, 25fps). There were 30 clips in total, sourced from 3 separate programmes originally broadcast by the British Broadcasting Corporation (BBC). Eye tracking was carried out using a tobii x50 eye tracker system, with the video displayed on a widescreen

plasma television, setup according to the ITU recommended guidelines [6]. This resulted in an eye gaze location, per participant, for every frame.

2.1. Eye Tracking Results and Analysis

Previous work [2] has shown that in open signed material there are two especially important regions - the signer's face and, to a lesser extent, the inset showing the original video.

As a tool to help verify results, the source video frames were manually classified, in 2 categories - 'signer signing?' (yes or no) and 'signer's facial orientation' (frontal, profile or unknown). In over 95% of the 7000 frames classified, when the face was frontal then the signer was signing. This suggests that the orientation of the signer's face is a good cue for whether or not signing is taking in any given frame.

We also investigate whether or not the orientation of the head is a good measure for whether or not the viewers are looking at the signer. Equation 1 shows how to calculate the proportion of eye gaze locations which are correctly predicted by the facial orientation of the signer, p_{correct}. Here a gaze location is 'correct' if it is on the signer when the face is frontal or is not on the signer when the face is profile.

$$p_{\text{correct}} = \frac{|F_{\text{pro}} \cap \overline{\mathcal{S}}| + |F_{\text{fro}} \cap \mathcal{S}|}{|\mathcal{S} \cup \overline{\mathcal{S}}|} \quad (1)$$

F_{fro} and F_{pro} are the sets of gaze locations when the signers face is frontal and profile respectively, and $\mathcal{S}$ is the set of gaze locations in which the viewer is fixated on the signer.

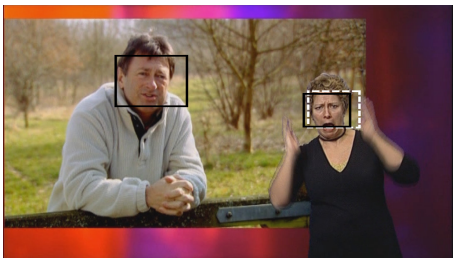
Using the previously described ground truth data, the proportion of accurately 'predicted' locations is found to be 80%. This implies that a predictor based on the orientation of the face of the signer would yield impressive results for gaze location prediction. Therefore a model for the orientation of the face of the signer is proposed in Section 3, which uses face detections and a tracker in an attempt to discover the orientation of the head of the signer. Although based on the orientation of the signer's head, this model is ultimately attempting to predict whether a viewer will be fixating on the signer or the inset for a given frame.

3. PROPOSED TECHNIQUE

3.1. Face Detection

The proposed gaze detector is based very heavily on face locations in a given frame for predicting the eye gaze position. We utilise the face detector proposed by Viola and Jones [7], which uses a cascade framework and Haar-like features to detect faces. The detector works on an individual frame basis, and detects faces of varying sizes based on a set of similar features. The detector is run with both frontal and profile feature cascades, so that 2 sets of detections are made. Each detection is specified by a box, within which a face has been detected. Figure 2 shows a sample frame and its associated face detections.

(a) Frame from a '*holby*' sequence, with correctly detected faces



(b) Frame from a '*meaning*' sequence, demonstrating an error with the face detection

Fig. 2: Output from the face detector with the solid line representing a frontal detection, and the hashed line a profile detection

Although the detector can make correct detections (Figure 2(a)), it can also make incorrect detections, such as the profile detection on the signer's face in figure 2(b). Therefore it is not possible to simply rely on the output of the face detector. This motivates the introduction of a tracker, which will decide how likely a face of a given orientation is at a given time.

3.2. Face Tracker

We use a grid based likelihood ratio tracker [8, pages 10.25 - 10.31] in order to detect not only the location of a face, but whether or not a face of a particular orientation (frontal or profile) exists at a given time. This will result in a surface which represents the likelihood ratio of a face existing at a given point versus no face being present in the frame at all.

We define states $s \in S$ as triples (s_x, s_y, s_d), which can be thought of as multiple grid-planes, each representing a tracking direction, $s_d \in D$. Here we choose $|D|$ to be 4, representing perpendicular directions in the 2-dimensional plane (up, down, left and right). Measurements $\xi_k \in \Xi$ represent the locations of detected faces in frame k as detected by the face detector from Section 3.1, and can be expressed as a coordinate pair $(\xi_{x,k}, \xi_{y,k})$. The likelihood ratio for frame k is written $\Lambda_k(s), \forall s \in S$. The likelihood ratio recursion begins with a motion update, shown in Equation 2.

$$\Lambda_k^-(s|s_d) = \Re(\Pi_k(s_d), s_d) * \frac{f_d}{|f_d|} \quad (2)$$

where $*$ represents a convolution, $\Lambda_k^-(s|s_d)$ is the motion update likelihood ratio for a given direction plane s_d and $\mathfrak{R}(H, v)$ rotates the plane H by $\frac{v\pi}{2}, v \in \{0, 1, 2, 3\}$. $\Pi_k(d)$ is the result of leaking likelihood ratio between directional planes, shown in Equation 3:

$$\Pi_k(d) = \sum_{i \in D} \Lambda_{k-1}(s|s_d = i) \cdot l(i, d) \tag{3}$$

$l(i, d)$ is the leak coefficient from direction plane i to plane d. In our implementation we choose a rotationally symmetric function such that $l(i, d) = 0.5 \cdot \delta_{\beta,0} + 0.1 \cdot \delta_{\beta,1}$, where $\delta_{i,j}$ is the kronecker delta and $\beta = |i - d| \pmod{|D|}$. $d, i \in \{0, 1, 2, 3\}$ in this work, since $|D| = 4$. f_d is a vector of filter coefficients defined in Equation 4:

$$f_d(i) = \begin{cases} 1 & 0 \leq i < c_d \\ e^{-\tau(i-c)} & c_d \leq i < n \end{cases} \tag{4}$$

$f_d(i)$ is the ith coefficient of the motion prediction filter of length n. The filter is designed such that pixel movements up to distance c_d are equally likely, after which the likelihood dies away exponentially. Values of $c_d = 3$ for horizontal direction planes, and $c_d = 1$ for the vertical planes were used, along with a decay factor of $\tau = 0.25$.

The next step of the recursion involves calculating the likelihood ratio $\mathfrak{L}_k(\xi|s_x, s_y)$ for frame k, measurement ξ, given the states defined by s_x and s_y (the measurements are independent of direction). Equation 5 shows how this is calculated.

$$\mathfrak{L}_k(\xi|s_x, s_y) = e^{-[(\xi_x - s_x)^2 + (\xi_y - s_y)^2]/2\sigma^2} \tag{5}$$

ξ is a detection at point (ξ_x, ξ_y), and σ^2 is the variance of the 2-dimensional Gaussian.

Finally the likelihood ratio surface is updated with the new information as detailed in Equation 6.

$$\ln \Lambda_k(s) = \ln \Lambda_k^-(s) + \ln \mathfrak{L}_k(\xi_k|s) \text{ for } s \in S \tag{6}$$

This recursive system is repeated through the frames of the video and a likelihood ratio surface evolves. Since we are interested in the likelihood ratio of a detection at any point in the image plane, we marginalise over the different directional planes to form a 2-dimensional likelihood ratio surface, $\tilde{\Lambda}_k(x, y)$, as shown in Equation 7.

$$\ln \tilde{\Lambda}_k(x, y) = \sum_{s_d \in D} \ln \Lambda_k(s|s_d) \tag{7}$$

3.3. Gaze Predictor

We propose a gaze predictor which is based on the face detections from Section 3.1 and the likelihood ratio tracker from Section 3.2. The tracker is used to decide whether or not the gaze is concentrated on the signer or the inset. Two separate trackers are run on a predefined bounding box around the signer's face simultaneously - one tracking frontal face views, the other profile. The output of each of these trackers is a likelihood ratio surface for each frame, $\tilde{\Lambda}_k^{\text{fro}}$ for the frontal tracker and $\tilde{\Lambda}_k^{\text{pro}}$ for the profile tracker for frame k. Comaparison of the maxima of these surfaces allows a prediction of whether or not the viewer will be looking at the signer or the inset. Equation 8 describes the prediction process, with d being a constant.

$$G_{\text{sign}}^k = \begin{cases} 1 & \text{if } \max(\ln \tilde{\Lambda}_k^{\text{fro}}) - \max(\ln \tilde{\Lambda}_k^{\text{pro}}) > d \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The gaze predictor specifies that if $G_{\text{sign}}^k = 1$, i.e. the signer is signing, then the gaze prediction will be at the location of the maximum of $\tilde{\Lambda}_k^{\text{fro}}$, i.e. the tracked location of the head of the signer. If $G_{\text{sign}}^k = 0$ then the viewer will be looking at the inset, and hence the predictor returns this region.

4. RESULTS

We use the eye tracking data collected as detailed in Section 2 to verify both the face tracker and gaze predictor.

4.1. Face Tracking Results

The face tracker is used to predict whether or not the viewers will be concentrating on the signer or elsewhere for any given frame. To this end we investigate the accuracy of G_{sign}^k. We define $G_{\text{sign}}^k = 1$ to be correct if the viewer is looking at the signer's face for frame k. From this we calculate recall, which is a measure of the proportion of false negatives recorded. In this scenario we are primarily concerned in minimising the false negatives, since it is more important to correctly predict fixations on the signer than it is to predict them on the inset. Equations 9 and 10 define precision and recall respectively, and Figure 3 shows how they variy with d (from Equation 8) for a set of 30 sequences.

$$\text{Precision} = \frac{\text{Number Correct}}{\text{Number Correct} + \text{False Positives}} \tag{9}$$

$$\text{Recall} = \frac{\text{Number Correct}}{\text{Number Correct} + \text{False Negatives}} \tag{10}$$

As can be seen in Figure 3 it is possible to attain a 95% recall value with $d = -5$. This is the value of d used in this work, as it yields high recall, but a suitably small number of false positive detections (the precision is over 80%).

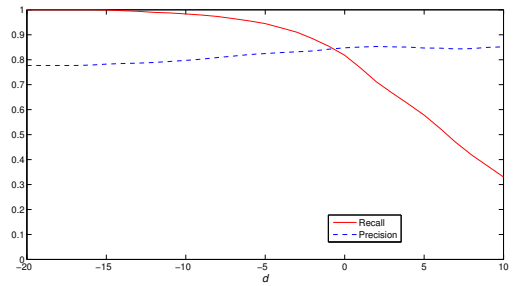
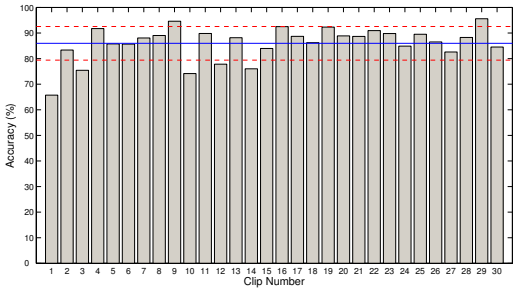


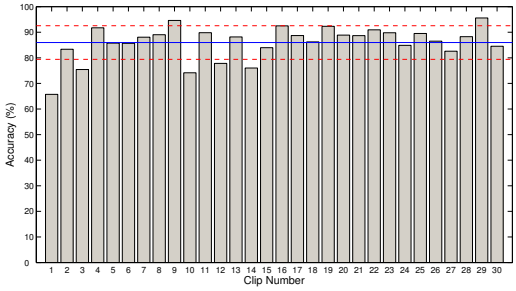
Fig. 3: Average Precision and Recall for '*meaning*' sequences, varying with d from Equation 8

4.2. Gaze Prediction Results

In order to assess the accuracy of the gaze predictor presented in Section 3.3 the results of the eye tracking study are compared to the gaze predictor. Since the gaze predictor returns boxes of high gaze likelihood, either the face of the signer or the inset, we say an eye track result has been correctly predicted if it falls within the predicted region for a given frame. From this we calculate a percentage of the eye tracking results that were correctly predicted by the gaze predictor.



(a) Accuracy results predicting gaze location as either signer or inset



(b) Accuracy results when localising the 'inset' predictions to face detections

Fig. 4: Percentage of eye track locations correctly predicted by the gaze tracker for 30 different clips. Including the mean and a standard deviation either side of the mean.

Figure 4(a) shows the individual results for 30 different video sequences. The results vary from 70% to 95%, with the mean being 86%. These results, however, only localise the gaze prediction to either the signer's face or the inset video. In order to attempt to further localise the gaze prediction within the inset we use the face detections found in Section 3.1 - defining an eye gaze location within the inset as predicted correctly if it is within the bounding box of a face detection. Figure 4(b) shows the results for the same 30 clips as before, which vary from 60% to 95%, with the mean being 79%.

5. CONCLUSIONS AND FUTURE WORK

In this paper a gaze prediction technique based on face detection and tracking has been described. A grid-based likelihood ratio tracker has been developed which will allow decisions regarding the orientation of a face (frontal or profile) to be made using detections from an error-prone detector. It has been shown that the face tracker can achieve a recall of over 95% when comparing the output with gaze locations from an eye tracking study. The gaze prediction routine has been shown to achieve an average of over 80% correct gaze locations, when compared with the same eye tracking study.

Improvements could be made to this system by introducing face tracking across the entire frame, as opposed to mainly around the signer, however this would result in a corresponding increase in computational complexity. Introducing the cues demonstrated in [2] could also see an improvement in performance. The gaze predictions can be used for a variety of applications, including variable quality coding and error protection. These applications require a subjective quality study to be undertaken to investigate the effects noticed by viewers.

6. REFERENCES

[1] L. J Muir and I. E. G Richardson, "Perception of sign language and its application to visual communications for deaf people," *J. Deaf Stud. Deaf Educ.*, vol. 10, no. 4, pp. 390–401, 2005.

[2] S. J. C Davies, D Agrafiotis, C. N Canagarajah, and D Bull, "Perceptually optimised coding of open signed video based on gaze prediction," *Electronics Letters*, vol. 39, no. 21, pp. 1135–1136, October 2007.

[3] D Agrafiotis, N Canagarajah, D Bull, and M Dye, "Perceptually optimised sign language video coding based on eye tracking analysis," *Electronics Letters*, vol. 39, no. 24, pp. 1703–1705, 27 Nov. 2003.

[4] D Argrafiotis, N Canagarajah, D Bull, H Twyford, J Kyle, and J Chung How, "Optimised sign language video coding based on eye-tracking analysis," *Visual Communications and Image Processing, Proc. of SPIE*, vol. 5150, pp. 1244–1252, 2003.

[5] L Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. on Image Processing*, vol. 13, no. 10, pp. 1304–1318, Oct 2004.

[6] I. T Union, "Recommendation itu-r bt.500-11: Methodology for the subjective assessment of the quality of television pictures," 2002.

[7] P Viola and M. J Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[8] D. L Hall and J Llinas, Eds., *Handbook of Multisensor Data Fusion*, CRC Press, 2001.

7. ACKNOWLEDGEMENTS

Thank you to Henry Knowles for his assistance with designing the face tracker. This work was funded by the British Broadcasting Corporation (BBC).