Davies, S. J. C., Agrafiotis, D., Canagarajah, C. N., & Bull, D. R. (2009). A multicue Bayesian state estimator for gaze prediction in open signed video. IEEE Transactions on Multimedia, 11(1), 39 - 48. 10.1109/TMM.2008.2008916

A Multicue Bayesian State Estimator for Gaze Prediction in Open Signed Video

Sam J. C. Davies, *Student Member, IEEE*, Dimitris Agrafiotis, *Member, IEEE*, C. Nishan Canagarajah, *Member, IEEE*, and David R. Bull, *Senior Member, IEEE*

***Abstract*—We propose a multicue gaze prediction framework for open signed video content, the benefits of which include coding gains without loss of perceived quality. We investigate which cues are relevant for gaze prediction and find that shot changes, facial orientation of the signer and face locations are the most useful. We then design a face orientation tracker based upon grid-based likelihood ratio trackers, using profile and frontal face detections. These cues are combined using a grid-based Bayesian state estimation algorithm to form a probability surface for each frame. We find that this gaze predictor outperforms a static gaze prediction and one based on face locations within the frame.**

***Index Terms*—Eye-tracking, face detection, gaze prediciotn, video coding.**

I. Introduction

THE ability to predict where viewers of a video are looking at any given time would be a useful tool in many video coding scenarios—from bit rate allocation to error protection. Studies have shown that eye gaze (i.e., the point on which a human's eyes are focusing) is both a top-down and bottom-up guided process [15]. Previous work on gaze prediction has generally been noncontext specific [16], [4] and therefore a largely bottom up process. Many techniques [5], [6], [18] involve the bottom-up concept of saliency [13], [17], which combines low level features of the video (such as color contrast, intensity, orientation, etc.) into a saliency map, and it is from this map that predictions of gaze location are made.

Different video contexts have different gaze patterns associated with them e.g., football and open sign language [2]. Due to the omission of any top-down processes, saliency has been shown to be inadequate for such video contexts (including open sign language [11]). Other techniques [8] introduce the idea of a top-down approach, but rely instead on categorizing objects within an image and then associating predetermined probabilities of eye-fixation with each object. Although this incorporates a top-down approach it ignores prior knowledge available about the scene.

Manuscript received May 21, 2008; revised September 08, 2008. Current version published January 08, 2009. The work of S. J. C. Davies was supported by the British Broadcasting Corporation (BBC). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shrikanth Narayanan.

The authors are with the Department of Electrical and Electronic Engineering, University of Bristol, Bristol, BS8 1UB, UK (e-mail: sam.davies@bristol.ac.uk; d.agrafiotis@bris.ac.uk; nishan.canagarajah@bris.ac.uk; dave.bull@bris.ac.uk).

Digital Object Identifier 10.1109/TMM.2008.2008916

Fig. 1. Sample frame from one of the "*holby*" open signed video sequences.

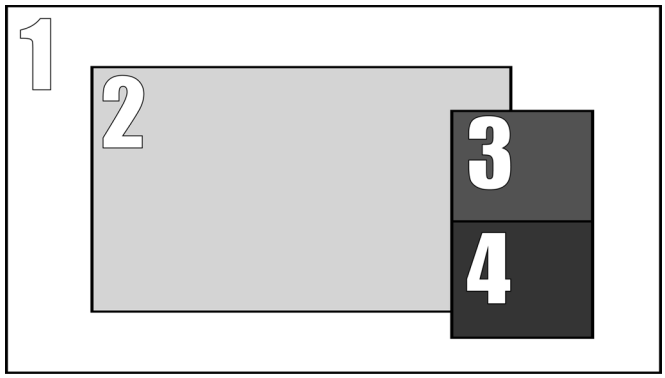

Fig. 2. Diagram of the different regions in an open signed frame: 1—background; 2—program inset; 3—signer's face; 4—signer's body.

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 concepts such as emotion. Fig. 1 shows a sample frame from open signed video footage, and Fig. 2 shows how an open signed frame can be divided up into distinct regions.

Open sign language is an excellent video context on which to perform gaze prediction due to the well-defined nature of the gaze patterns as shown in [14], [11], [1] and [3]. Previous work has been reported on analyzing the gaze patterns of sign language in various forms, including videoconferencing [3] and open sign video [11], with the aim of coding video with perceptually varying quality. Ciaramello and Hemami [9] developed an objective metric for measuring the intelligibility of sign language before using it to optimize rate distortion in coding video [10]. These studies all showed that within the sign language context it is possible to make coding gains without loss of intelligibility or perceived quality.

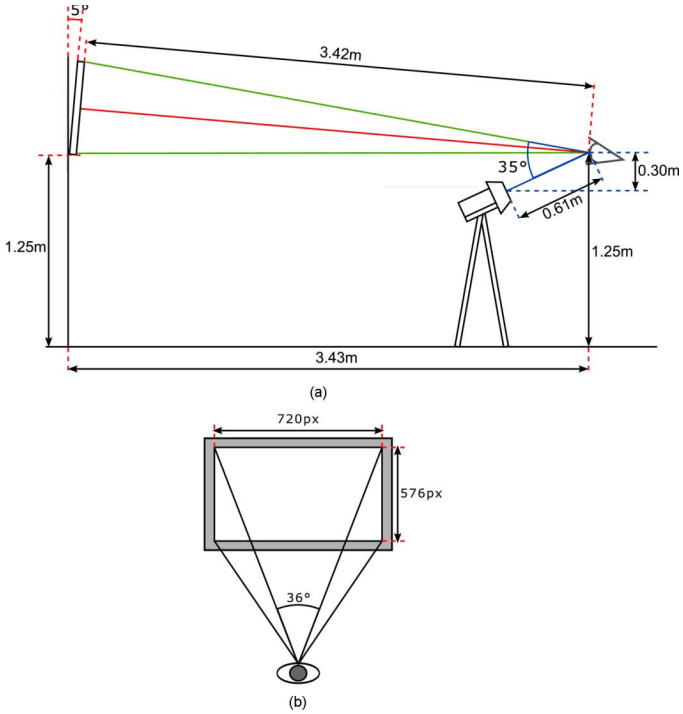

Fig. 3. Eye tracking experiment setup. (a) Side on view. (b) Plan view.

In order to design an open sign language gaze tracker, we first perform an eye tracking study (see Section II) which identifies several cues which might be of use to predict the location of a viewer's gaze. These cues include motion of the signer's hands, the location of faces within the frame, the orientation of the signer's head, and shot changes within the inset. We then develop methods to extract these cues from the video, including designing a grid-based likelihood ratio tracking system to predict the orientation of the signer's face. A grid-based Bayesian state estimation technique is employed to combine the cues into a probability surface which can be used to predict the viewer's gaze.

This paper is organized as follows: Section II describes the eye tracking study carried out with sign language material, before Section III investigates the suitability of possible cues for a gaze prediction routine. Sections IV and V detail how two of these cues (Shot Changes and Face Orientation respectively) can be generated from the raw video, including in Section V-B a face tracking routine based on face detections. We then use these cues to develop some gaze prediction routines in Section VI, initially a single cue predictor (Section VI-A) before moving onto a multicue model (Section VI-B). The paper is concluded in Section VII, including a discussion on further work.

II. 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, 25 fps). There were 30 clips in total with durations between 20 and 45 s, sourced from three separate programs originally broadcast by the British Broadcasting Corporation (BBC), here referred to as *holby, meaning* and *caribbean*. The clips were recorded off-air, as MPEG-2 broadcast streams, with a bitrate of around 5 Mb/s.

Fig. 4. Sample frame with eye tracking fixation density plotted as hotspots. The map is generated by overlaying all of the subjects gaze patterns for a sequence over one chosen frame.

Eye tracking was carried out using a tobii x50 eye tracker system, which consists of a control PC and a completely non-intrusive eye tracking device. The eye tracker uses an infra red camera to track the reflections of the pupils of the subject, and via a calibration routine the computer calculates the gaze location. The system operates at 50 Hz, so the mean of each pair of samples is taken to yield a single gaze location per frame.

The experiments were performed in a room with dimmed lighting, with the participants sitting in front of a 52 inch (16:9) plasma screen (111 cm $\times$ 62 cm active visible area) with the resolution set at 720×576 (Standard Definition). In order to comply with the Preferred Viewing Distance (PVD) outlined in the ITU Recommendation [19] the subject sat approximately 3.5 m from the screen, as shown in 3. The eye tracker is accurate within the nearest 0.5 visual degrees, which corresponds to an accuracy on the screen of 2.98 cm. This allows a gaze prediction accuracy of around 20 pixels horizontally and 28 pixels vertically.

Fig. 4 is a representation of the spatial distribution of the eye-tracked gaze locations and from this we see that clearly the most significant region is the signer's head. This supports previous work in both videoconferencing contexts [3] and open sign contexts [11]. The program inset video also has gaze fixations localized on it, although Fig. 4 provides no information as to when these occur. On inspection of the video we notice that inset fixations tend to occur on shot changes and pauses in signing and when these saccades occur, if there are faces within the inset video then the fixations are often localized at these positions.

III. Possible Gaze Prediction Cues

A. Shot Changes

When a shot change occurs in the inset it is natural for the viewer to exhibit a brief saccade from the signer to the inset in order to analyze the new scene. The true locations of the shot changes in the inset were found manually. We then investigated the proportion of these shot changes which are followed by a gaze location which is not defined as part of the signer's face, within a given time frame.

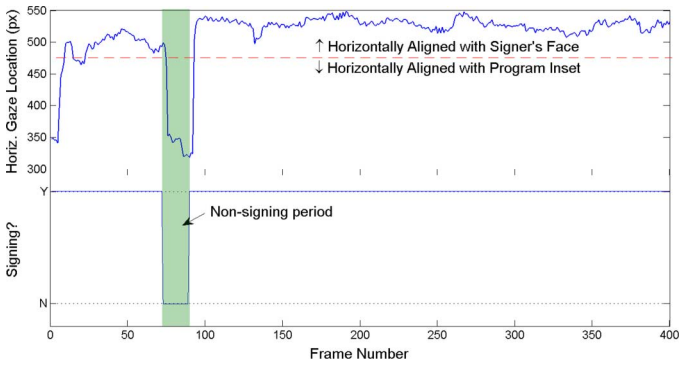

Fig. 5. Comparison of horizontal gaze position with signing periods.

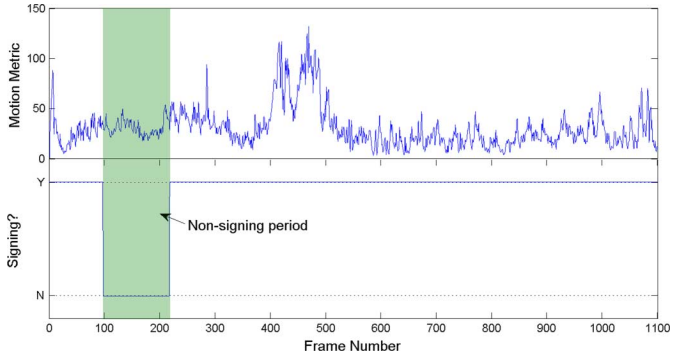

Fig. 6. Comparison of signing periods with a simple metric for the motion of the signer's hands.

For the *meaning* clips it was found that 68% of program inset shot changes were followed by a saccade away from the signer's face within 25 frames (1 s).

It is then proposed that shot changes in the inset program video will form a good cue for predicting the gaze position between the signer's face and the inset video.

B. Signer Signing

Conceptually it seems intuitive that when the signer is signing, the attention of the viewer will be concentrated around the signer, and conversely when there is no signing taking place the gaze will move to the inset. In order to demonstrate this hypothesis, each of the source video frames was manually classified as "signer signing?" (yes or no). The video clips used are representative of the open signed output offered by the British Broadcasting Corporation (BBC) and the ground truth results show that signing is taking place in 87% of the frames.

Fig. 5 shows the horizontal gaze position for an observer of a sequence (*holby_03*) and a plot of whether a signer is signing for any given frame, obtained from ground truth data. These results imply that there is a high correlation between periods of signing and gaze location (signer or program inset).

In order to attempt to quantify this relationship between signing and gaze location we define $\mathcal{GT}_{\text{sign}}$ as the set of frames labeled as "signing" (by manual ground truthing) and $\mathcal{ET}_{\text{sign}}$ as the set of frames in which the eye gaze locations are concentrated on the signer's face (defined as being within a predefined bounding box). We select the face as the region of interest since previous work ([11], [1]) has shown that fluent sign language users fixate almost entirely around the mouth when watching sign language. From these we define two measurements outlined in (1) and (2), the former measuring the proportion of frames where the gaze is on the signer which are labeled as signing, and the latter the proportion of the "signing" frames which have gaze locations on the signer.

$$p_{\text{sign}}^{\mathcal{ET}} = \frac{|\mathcal{GT}_{\text{sign}} \cap \mathcal{ET}_{\text{sign}}|}{|\mathcal{ET}_{\text{sign}}|} \tag{1}$$

$$p_{\text{sign}}^{\mathcal{GT}} = \frac{|\mathcal{GT}_{\text{sign}} \cap \mathcal{ET}_{\text{sign}}|}{|\mathcal{GT}_{\text{sign}}|}. \tag{2}$$

For the 7000 ground truthed frames, and five study participants, values of $p_{\text{sign}}^{\mathcal{ET}} = 0.95$ and $p_{\text{sign}}^{\mathcal{GT}} = 0.82$ were obtained. This implies that 95% of the time the observer is looking at the signer, signing is taking place, and that for 82% of the time signing is taking place the observers are concentrating on the signer. Therefore being able to detect whether signing is occurring for any given frame will be a good cue for predicting the gaze locality. The following possibilities were considered.

1) Motion of the Hands: The motion of the signer's hands intuitively appears to be a good cue for detecting whether or not a signer is signing at any given time.

The sequence is coded using the h.264 reference coder, with an IPPP structure. The motion prediction for the P-frames is limited to the previous frame and on a macroblock basis only. This yields one motion vector per macroblock per frame. The motion metric MS is defined as follows:

$$MS = \sum_{\{i,j\} \in R} \|mv_{i,j}\| \tag{3}$$

where $mv_{i,j}$ is the h.264 motion vector for the (i,j)th macroblock and R is the set of all (i,j) pairs in region 4 of Fig. 2.

Fig. 6 shows this motion metric for sequence *meaning_06*, along with the signing frames, defined from the ground truth data.

The nonsigning period is during a section of low motion compared to the global mean, and just before signing starts again there is a sudden increase in motion—these two facts could be used to help define a signing estimator from the motion in the region around the signer's hands. However, these two cues (low motion and sudden increase) are not exclusive to this nonsigning region and occur elsewhere within this clip—making it difficult to localize nonsigning periods.

This motion metric could be improved by segmenting out the hands of the signer and looking at the entropy of the direction of the motion (towards the end of a signing sequence, a signer will lower their hands to a stop). However initial investigations show that other, simpler cues can provide superior information.

2) Orientation of the Head: Observation of open-signed material leads to the conclusion that the orientation of the signer's head might correlate well with whether or not they are signing, since when not signing the signer tends to look towards a preview monitor placed virtually to imply they are watching the inset program. This, therefore, could in turn be a good cue for predicting the gaze location between the signer's face and the inset program.

The source video frames were again manually classified according to "signer's facial orientation" (frontal, profile or unknown). In over 95% of the frames classified as "signer's face

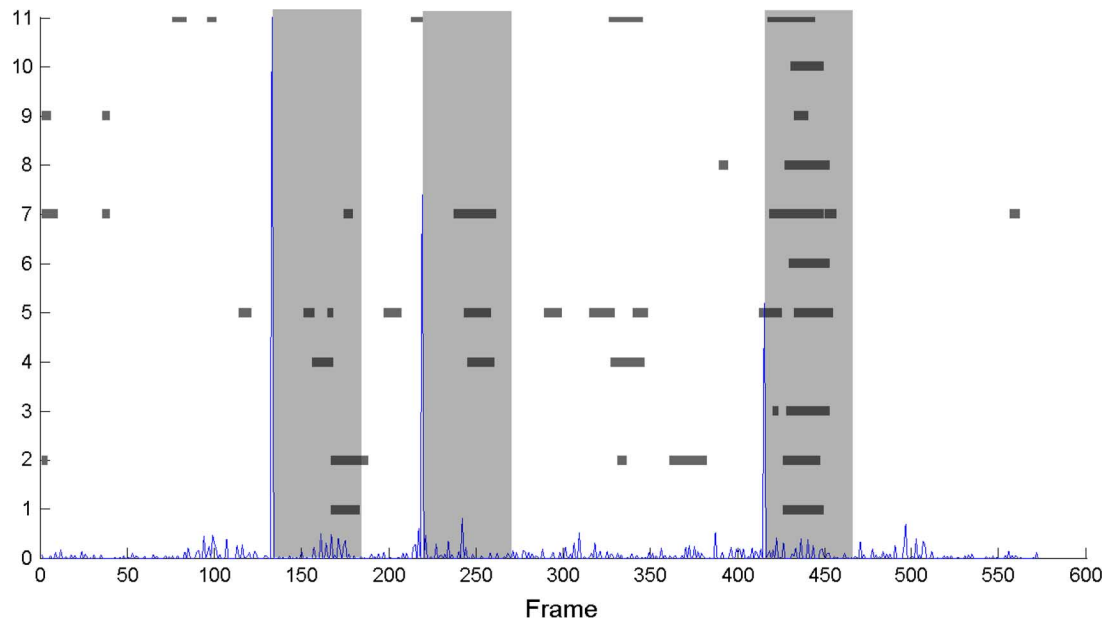

Fig. 7. Shot change metric (d_2), with 50 frame windows after shot changes highlighted in light grey. Each row of the horizontal dark grey bars represents an observer fixating on the inset.

frontal" the signer was signing. In contrast, in only 0.15% of the frames was the signer not signing whilst their face was frontal. This suggests that the orientation of the signer's face is a good cue for whether or not signing is taking place in any given frame. These statistics are consistent with the etiquette employed in open signing, where a monitor is positioned such that when no signing is necessary it gives the impression that the signer is watching the program itself.

We also investigated whether or not the orientation of the head is a good measure for whether or not the viewers are looking at the signer. Equation (4) calculates the proportion of eye gaze locations which are correctly predicted by the facial orientation of the signer, $p_{\text{head}}^{\text{cor}}$. 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:

$$p_{\text{head}}^{\text{cor}} = \frac{|\mathcal{GT}_{\text{profile}} \cap \overline{\mathcal{GT}_{\text{sign}}}| + |\mathcal{GT}_{\text{frontal}} \cap \mathcal{GT}_{\text{sign}}|}{|\mathcal{ET}_{\text{sign}} \cup \overline{\mathcal{ET}_{\text{sign}}}|}. \quad (4)$$

$\mathcal{GT}_{\text{frontal}}$ and $\mathcal{GT}_{\text{profile}}$ are the sets of gaze locations when the signers face is frontal and profile respectively, and $\mathcal{ET}_{\text{sign}}$ is the set of gaze locations in which the viewer is fixated on the signer (as previously).

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.

Cherniavsky *et al.* [7] propose the usage of frame differencing and support vector machines to address the same "signer-signing" problem. This technique yields encouraging results for signer-only video but in the open signed scenario problems such as the inset video being part of the background of the signer cause various problems. Due to the fact that the signers adhere to simple conventions regarding when to face the camera (when not signing they appear to watch the inset video), accurately estimating the orientation of the head is a more appropriate cue for this context of signed video.

Therefore a model for the orientation of the face of the signer is proposed in Section V, 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.

IV. Shot Change Detection

To detect shot changes in the inset program video a simple technique is employed. A 32-bin histogram of the intensity of the predetermined inset region (see Fig. 2, region 2) for each frame and the L_2 norm of the difference between consecutive frames is calculated:

$$\Delta_i = \|h_{i+1} - h_i\| = \sqrt{\sum_j (h_{i+1,j} - h_{i,j})^2} \quad (5)$$

where h_i is the histogram vector at time $i, h_{i,j}$ is the jth bin of the histogram at time i, and Δ_i is the consecutive difference between histograms at time i.

Shot changes are most likely to be related to the 2nd difference of this motion measure—i.e., a sudden large change in the difference measure (L_2) will represent a shot change. Fig. 7 shows this normalized shot change detection metric (d_2) and the fixations on the inset for each of the test subjects. The 50 frame windows after each shot change cover fixations, which supports the idea that the shot changes cause fixations on the inset.

Using a simple threshold of ten times the mean the metric to detect shot changes within a sequence yields an average precision of 80% and an average recall of 82% (a detection was labeled as correct if it was within ten frames of a manually labeled shot change).

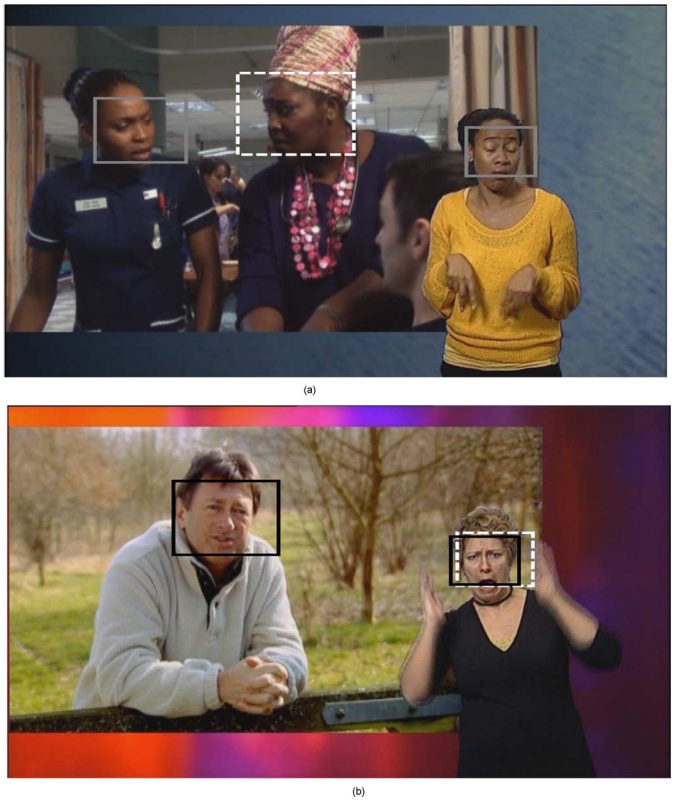

Fig. 8. Output from the face detector with the solid line representing a frontal detection, and the hashed line a profile detection. (a) Frame from a "*holby*" sequence, with correctly detected faces. (b) Frame from a "meaning" sequence, demonstrating an error with the face detection.

V. Face Tracker

A. Face Detection

The proposed gaze detector is based very heavily on face locations in a given frame for predicting the eye gaze position. We utilize the face detector proposed by Viola and Jones [20], 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 two sets of detections are made. Each detection is specified by a box, within which a face has been detected. Fig. 8 shows a sample frame and its associated face detections.

Although the detector can make correct detections [Fig. 8(a)], it can also make incorrect detections, such as the profile detection on the signer's face in Fig. 8(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.

B. Tracking

We use a grid based likelihood ratio tracker [12] in order to detect not only the location of a face, but also its orientation (frontal or profile) at a given time. This results 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-D plane (up, down, left and right). Measurements $\xi_k \in \Xi$ represent the centroids of detected faces in frame k as detected by the face detector from Section V.A, 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 (6).

$$\Lambda_k^-(s \mid s_d) = \Re(\Pi_k(s_d), s_d) * \frac{f_d}{|f_d|} \tag{6}$$

where $*$ represents a convolution, $\Lambda_k^-(s|s_d)$ is the motion update likelihood ratio for a given direction plane s_d and $\Re(H, v)$ rotates the plane H by $(v\pi)/(2), v \in \{0, 1, 2, 3\}$. $\Pi_k(d)$ is the result of leaking likelihood ratio between directional planes, shown in (7):

$$\Pi_k(d) = \sum_{i \in D} \Lambda_{k-1}(s \mid s_d = i) \cdot l(i, d). \tag{7}$$

$l(i, d)$ 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 $\beta = |i - d|(\mathrm{mod}|D|)$. $d, i \in \{0, 1, 2, 3\}$ in this work, since $|D| = 4$. f_d is a vector of filter coefficients defined in (8):

$$f_d(i) = \begin{cases} 1 & 0 \le i < c_d \\ e^{-\tau(i-c)} & c_d \le i < n. \end{cases} \tag{8}$$

$f_d(i)$ 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 $\mathfrak{L}_k(\xi \mid s_x, s_y)$ for frame k, measurement ξ, given the states defined by s_x and s_y (the measurements are independent of direction). Equation (9) shows how this is calculated:

$$\mathfrak{L}_k(\xi \mid s_x, s_y) = e^{-[(\xi_x - s_x)^2 + (\xi_y - s_y)^2]/2\sigma^2}. \tag{9}$$

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

Finally the likelihood ratio surface is updated with the new information as detailed in (10):

$$\ln \Lambda_k(s) = \ln \Lambda_k^-(s) + \ln \mathfrak{L}_k(\xi_k|s) \text{ for } s \in S. \tag{10}$$

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 marginalize over the different directional planes to form a 2-D likelihood ratio surface, $\tilde{\Lambda}_k(x, y)$, as shown in (11):

$$\ln \tilde{\Lambda}_k(x, y) = \sum_{s_d \in D} \ln \Lambda_k(s \mid s_d). \tag{11}$$

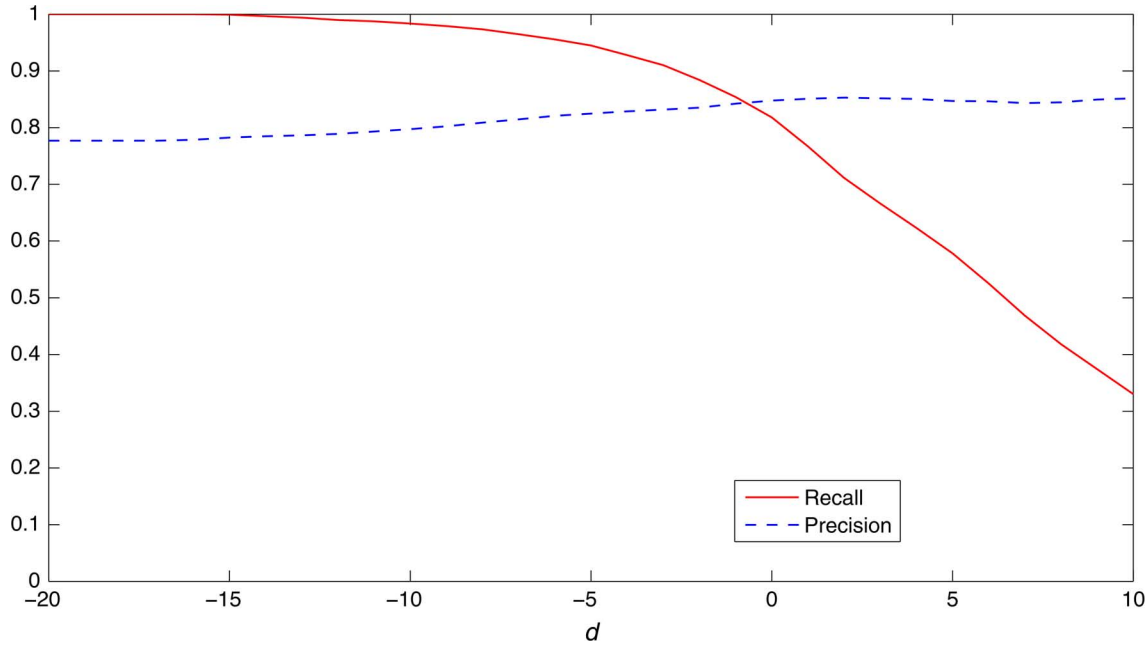

Fig. 9. Average Precision and Recall for "*meaning*" sequences, varying with d from (14).

C. 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^k_{sign}. We define $G^k_{\text{sign}} = 1$ 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 minimizing 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 (12) and (13) define precision and recall respectively, and Fig. 9 shows how recall varies with d [from (14)] for a set of 30 sequences.

$$\text{Precision} = \frac{\text{Number Correct}}{\text{Number Correct} + \text{False Positives}} \tag{12}$$

$$\text{Recall} = \frac{\text{Number Correct}}{\text{Number Correct} + \text{False Negatives}}. \tag{13}$$

As can be seen in Fig. 9 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%).

VI. Gaze Predictor

A. Single Cue Gaze Predictor

We propose a gaze predictor which is based on the face detections from Section V.A and the likelihood ratio tracker from Section V-B. 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, $\tilde{\Lambda}^{\text{fro}}_k$ for the frontal tracker and $\tilde{\Lambda}^{\text{pro}}_k$ for the profile tracker for frame k. Comparison 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 (14) describes the prediction process, with d being a constant:

$$G^k_{\text{sign}} = \begin{cases} 1, & \text{if } \max\left(\ln \tilde{\Lambda}^{\text{fro}}_k\right) - \max\left(\ln \tilde{\Lambda}^{\text{pro}}_k\right) > d \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

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

In order to assess the accuracy of the gaze predictor presented here, the results are compared with those from the eye tracking study. 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.

Fig. 10(a) shows the individual results for 30 different video sequences. The results vary from 70% to 95%, with the mean being 86% and the standard deviation 0.066. These results, however, only localize the gaze prediction to either the signer's face or the inset video. In order to attempt to further localize the gaze prediction within the inset we use the face detections found in Section V-A—defining an eye gaze location within the inset as predicted correctly if it is within the bounding box of a face detection. Fig. 10(b) shows the results for the same 30 clips as before, which vary from 60% to 95%, with the mean being 79%, and the standard deviation 0.085. As expected, the more specific the localization the less accurate the prediction and the more varied the results.

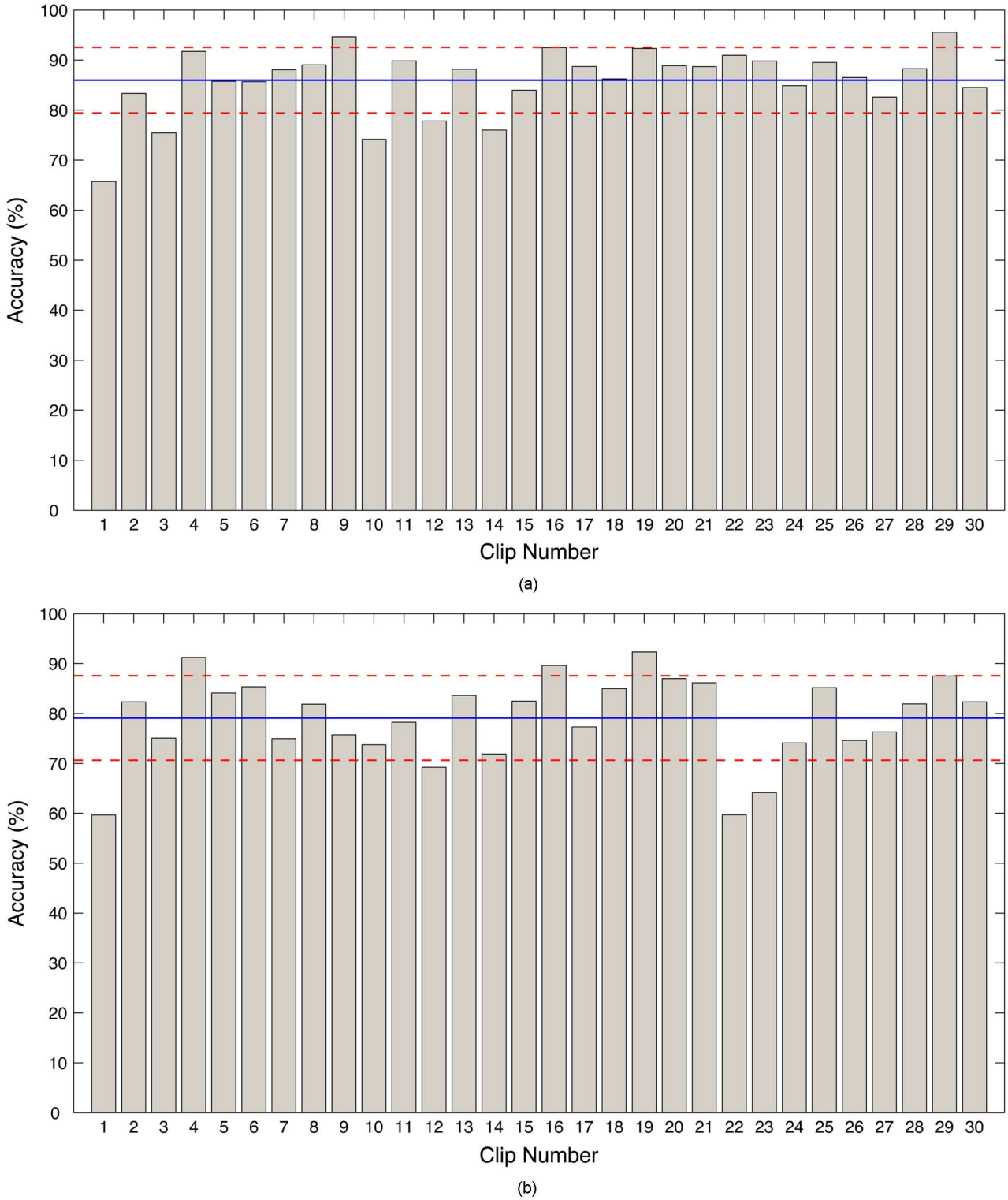

Fig. 10. Percentage of eye track locations correctly predicted by the gaze tracker for 30 different clips. Including the mean and a standard deviation either side of the mean. (a) Accuracy results predicting gaze location as either signer or inset. (b) Accuracy results when localizing the "inset" predictions to face detections.

Clips 22–24 have a significant drop in accuracy in the face-localized technique shown in Fig. 10(b) compared to that shown in Fig. 10(a). This is due to the nature of the content of the video-clips 21–30 are taken from a nature program and therefore there are no human faces in the original video. Therefore when the face localization is invoked, there are no faces to find in the inset and therefore the gaze predictions are essentially random across the inset.

B. Multiple Cue Gaze Predictor

We use grid-based Bayesian state estimation to combine the chosen cues (detected face locations, signer's facial orientation

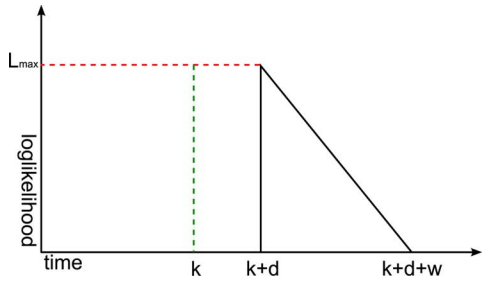

Fig. 11. Loglikelihood increases over time for detected signing status changes and shot changes.

and shot change detections). This means that each pixel has an associated likelihood of being a fixation position for any given frame, and the likelihood surface for the following frame is predicted from the current surface and the updated cues.

Each cue will define a region which it associates with a high likelihood of gaze location for each frame. We use these regions to increase the overall likelihood map. The area of the frame which has not been predicted to contain gaze locations by any of the predictors has its likelihood reduced. The likelihood of the subsequent frame is then predicted from this updated map and the process iterates.

We work in the loglikelihood domain for ease of calculation and at each update stage ensure that the range is restricted to $[-2, 5]$. In the following explanation of the technique all of the values used here were found empirically to optimize the output of the predictor compared to eye tracking results. As a prior distribution we have a uniform base loglikelihood across the frame with a slightly higher value in the signer's bounding box and the inset (here these were 0 and 3 respectively).

The algorithm has two parts: update and prediction. The prediction stage takes the form of a 2-D Gaussian filter across the entire frame—this is representing the principle that a position of high gaze likelihood for one frame will also be a likely gaze position for the following frame, but that there might be some slight variation in local position. Here we use a symmetrical 20×20 tap filter, with a standard deviation of 15.

The update stage consists of three separate procedures, one for each of the selected cues. In order to lower the likelihood in regions not labeled as high importance by any of the detectors, we lower the loglikelihood of the entire frame (here chosen empirically to be 0.05) and then add loglikelihood to the regions defined by the cues. Each of the cues defines regions and increments the loglikelihood accordingly:

- **Face Locations** The face detector returns face locations and these generate regions which are of high gaze probability. We combine this cue with the signer's face orientation to weight the face locations accordingly. If the detected face appears to be the signer's face when the signer is signing (frontal face tracked) or the detected face is in the inset when the signer is not signing (profile face tracked) then the detection is said to be "active" and a certain loglikelihood is added to the detected regions (here chosen to be 0.25). Conversely, if a detection is labeled as "inactive" (not signing and on signer, or signing and on inset) then a smaller loglikelihood is added to the detected facial region (here, 0.1).
- **Signer's Facial Orientation** The output of the signer's facial orientation tracker is an indication of whether or not the signer is signing at any given time (see Section V-C). We use this information to add loglikelihood uniformly across the inset when the signer is not signing, since this event results in saccades from the signer to the inset. Fig. 11 shows the form that this addition takes. k represents the point at which the signer stops signing, $L_{\max}$ the maximum loglikelihood added to the inset, d the delay length and w the length of the window over which increases occur (the latter three were set at 0.4, 0 and 10 respectively).
- **Shot Changes** Shot changes in the inset are detected and modulate the loglikelihood of the inset in the same way as the signer's facial orientation does. Fig. 11 details the way the values of the inset are increased when a shot change is detected at frame k. Here the delay (d) was set at 10, the window length (w) at 15 and the maximum increase $(L_{\max})$ at 0.4.

This completes the update stage and the prediction/update procedure is repeated for all subsequent frames. To generate the likelihood surface, the exponent of the value of the loglikelihood is taken for each pixel, and this new surface is normalized so that it sums to 1.

Since this grid-based Bayesian approach produces a likelihood surface for each frame rather than a predicted location, direct comparison is impossible. Therefore we use a threshold-based method of estimating gaze predictions from the likelihood surface. Fig. 12 shows the accuracy of the multicue gaze predictor, with a varying threshold. The threshold represents the proportion of each frame which is accepted as a gaze prediction (region of interest).

Fig. 13 shows the proportion of correct eye gaze locations for the 30 individual clips, with the ROI threshold set at 20% of the entire frame. The mean proportion of correctly predicted eye-track locations across all the test sequences is 90%.

It can be seen from Fig. 12 that the accuracy clearly varies with the proportion of each frame labeled as a Region of Interest (ROI). For an ROI coverage of 15%, an average accuracy of 87% is obtained, whilst with a coverage of 20% this accuracy is increased to 90%, with a standard deviation of 0.044. In comparison, a prediction model which estimates that the viewer will be continuously fixated on the signer's face has an average accuracy of 73%, whilst the single-cue techniques yield an 87% accuracy for inset-versus-signer (see Fig. 10(a)) and an 81% accuracy for localized faces (see Fig. 10(b)), where each technique has an ROI coverage area of 15–20%. Removing the shot detection cue from the multicue predictor results in decreases in prediction accuracies, varying across the selection of clips (2% to 8%). Therefore the multicue approach yields not only more accurate results than both the single-cue techniques and the static prediction, but also reduces the variance of the prediction accuracy.

The multicue technique therefore not only outperforms the single-cue gaze predictor but also provides more information in the form of a likelihood surface as opposed to discrete gaze prediction locations. This surface provides a relative importance

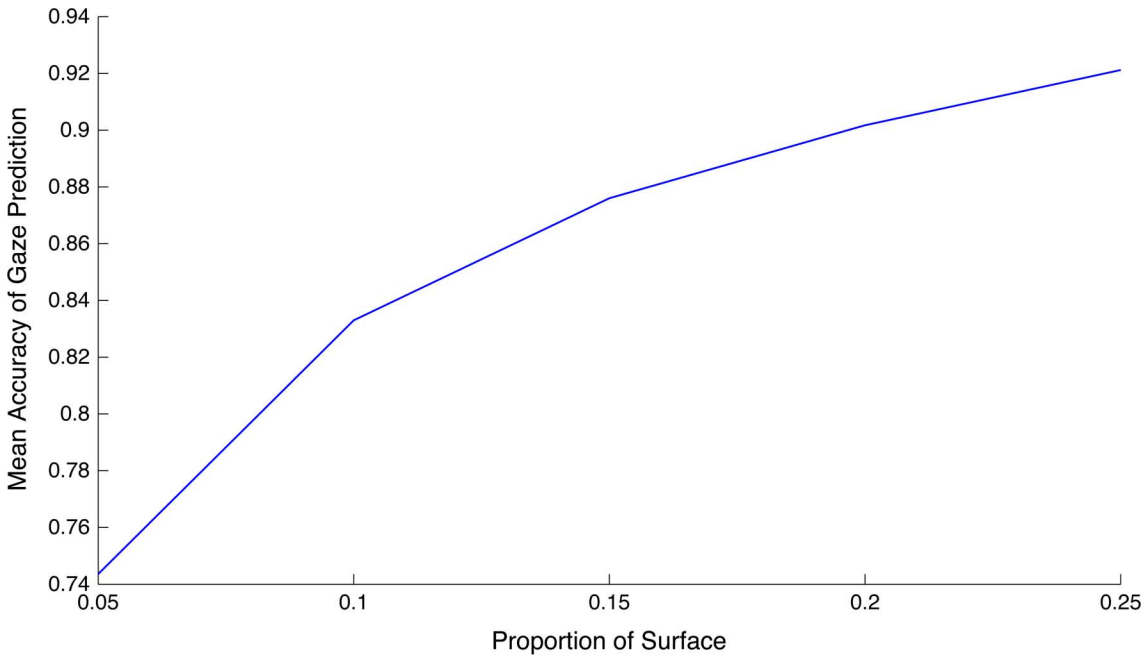

Fig. 12. Accuracy of the multicue gaze predictor, based on a varying threshold of the proportion of each frame accepted as a prediction location.

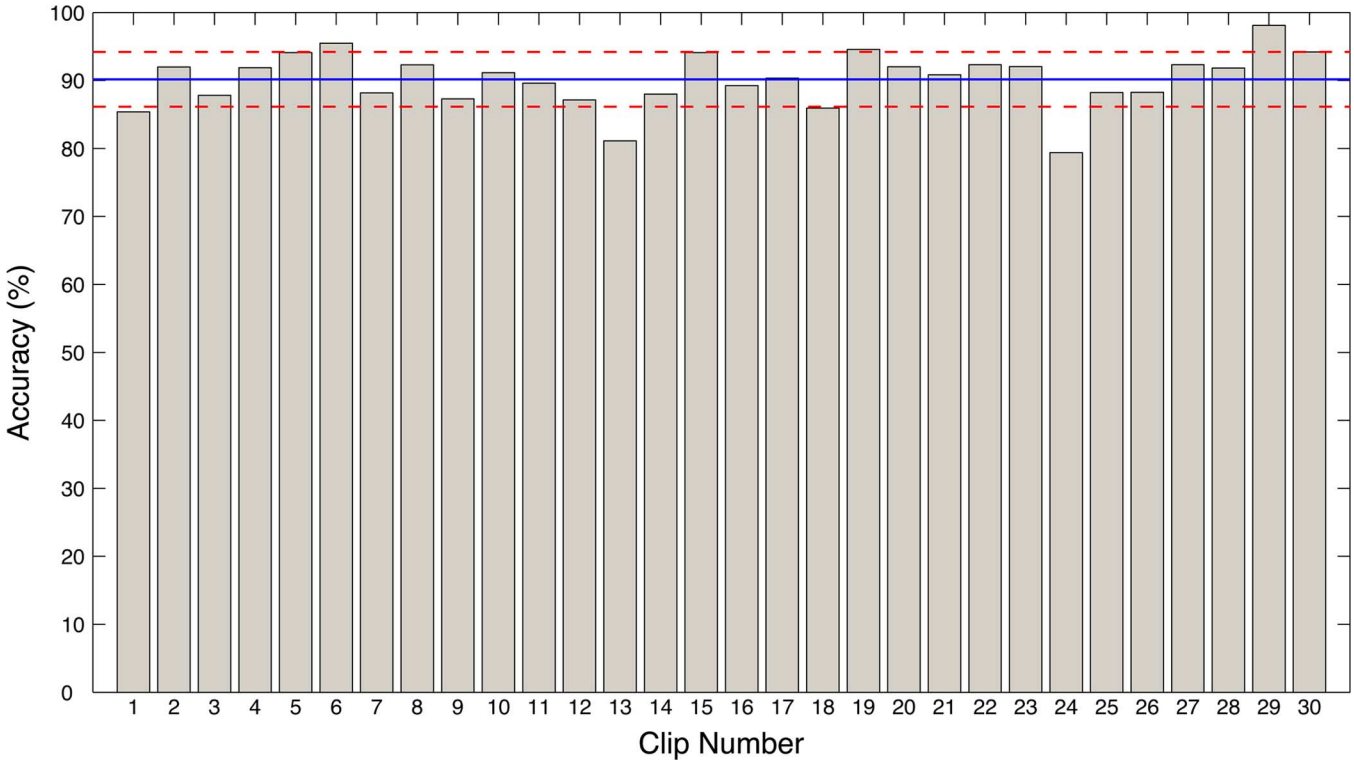

Fig. 13. Percentage of eye track locations correctly predicted by the multicue gaze predictor with the ROI threshold set at 20%. Includes mean and a standard deviation either side of the mean.

score across each frame-not just a binary ROI/nROI map, and this surface can be incorporated into the same end uses as the binary map can (e.g., variable quality coding).

VII. Conclusion

In this paper we have proposed a gaze prediction system for a specific video context, namely open sign language. This system uses a grid-based Bayesian state estimator to combine various cues, which themselves have been generated from the raw video. We have also investigated the validity of different cues (shot change in the inset program video, location of faces across the frame, the orientation of the signer's face and the motion of the signer's hands), and developed techniques for extracting these cues from the video at a signal level, including a novel use of a likelihood ratio tracker to detect the orientation of a face.

The proposed gaze prediction model produces a probability surface for each frame-showing the likelihood of a viewer looking at any point within the frame. The output of this was analyzed against eye tracking data for a large selection of clips and was shown to outperform both a uniform probability surface, and a surface generated using face locations within the frame.

This probability surface can be used for various different purposes within coding, including bit-rate allocation creating a perceptually optimized video codec, and error protection. This gaze prediction framework can be adapted to suit different video contexts using prior knowledge of the gaze pattern.

REFERENCES

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

[2] D. Agrafiotis, S. J. C. Davies, N. Canagarajah, and D. R. Bull, "Towards efficient context-specific video coding based on gaze-tracking analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, no. 4, pp. 1–15, 2007.

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

[4] E. Barth, J. Drewes, and T. Martinez, "Dynamic predictions of tracked gaze," in *Proc. 7th Int. Symp. Signal Processing and Its Applications*, July 2003, vol. 1, pp. 245–248.

[5] G. Boccignone and M. Ferraro, "Modelling gaze shift as a constrained random walk," *Phys. A*, vol. 331, pp. 207–218, 2004.

[6] W. Cheng, W. Chu, and J. Wu, "A visual attention based region-of-interest determination framework for video sequences," *IEICE Trans. Inform. Syst.*, vol. E88-D, no. 7, pp. 1578–1586, 2005.

[7] N. Cherniavsky, A. C. Cavender, R. E. Ladner, and E. A. Riskin, "Variable frame rate for low power mobile sign language communication," in *Assets'07: Proc. 9th Int. ACM SIGACCESS Conf. Computers and Accessibility*, New York, 2007, pp. 163–170, ACM.

[8] D. A. Chernyak and L. W. Stark, "Top-down guided eye movements," *IEEE Trans. Syst., Man, Cybern. B: Cybern.*, vol. 31, no. 4, pp. 541–522, Aug. 2001.

[9] F. M. Ciaramello and S. S. Hemami, "Can you see me now?," *Proc. SPIE: An Objective Metric for Predicting Intelligibility of Compressed American Sign Language Video*, vol. 6492, p. 64920M, 2007.

[10] F. M. Ciaramello and S. S. Hemami, "Complexity constrained rate-distortion optimization of sign language video using an objective intelligibility metric," *Vis. Commun. Image Process.*, vol. 6822, no. 1, p. 682213, 2008.

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

[12] , D. L. Hall and J. Llinas, Eds., *Handbook of Multisensor Data Fusion*. Boca Raton, FL: CRC, 2001.

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

[14] 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.

[15] E. Niebur and C. Koch, *The Attentive Brain*. Cambridge, MA: MIT Press, 1998, ch. chapter Computational Architectures for Attention, pp. 163–186.

[16] C. Privitera and L. W. Stark, "Algorithms for defining visual region-of-interest: Comparison with eye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, Sep. 2000.

[17] C. M. Privitera and L. W. Stark, *Neurobiology of Attention*. Amsterdam, The Netherlands: Elsevier, 2005, ch. chapter Scanpath Theory, Attention, and Image Processing Algorithms for Predicting Human Eye Fixations, pp. 24–28.

[18] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "Foveated analysis and selection of visual fixation in natural scenes," in *IEEE Int. Conf. Image Processing*, 2006, pp. 453–456.

[19] *Int. Telecommun. Union*, , 2002, Recommendation itu-r bt.500-11: Methodology for the subjective assessment of the quality of television pictures.

[20] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

Sam J. C. Davies (S'08) received the B.A. (Hons.) degree in mathematics from the University of Cambridge, U.K., and the M.Sc. (Distinction) degree in communications systems and signal processing from the University of Bristol. He is currently pursuing the Ph.D. degree at the Centre of Communications Research at the University of Bristol. His research interests include image and video coding, human perception of video and eye-tracking.

Dimitris Agrafiotis (M'01) received the M.Sc. (Distinction) degree in electronic engineering from Cardiff University, U.K., in 1998 and a Ph.D. from the University of Bristol, U.K., in 2003.

He is currently an RCUK Academic Fellow within the Centre for Communications Research at the University of Bristol. His research interests include image and video coding, video transmission over wireless networks, error resilience, gaze-tracking and related applications. He has published over 40 papers in these areas.

C. Nishan Canagarajah (M'95) received the B.A. (Hons.) and Ph.D. degrees in DSP techniques for speech enhancement, both from the University of Cambridge, Cambridge, U.K.

He is currently a Professor of Multimedia Signal Processing at the University of Bristol, U.K.. He was previously an Research Assistant and Lecturer at Bristol, investigating DSP aspects of mobile radio receivers. His research interests include image and video coding, image segmentation, content based video retrieval, 3–D video and image fusion. He is widely supported in these areas by industry, EU and the EPSRC. He has been involved in a number of EU FP5 and FP6 projects where the team has been developing novel image/video processing algorithms. He has published more than 160 papers and two books.

Prof. Canagarajah is a member of the EPSRC Peer Review College.

David R. Bull (M'93–SM'07) is Professor of signal processing and Head of the Signal Processing Group in the Electrical and Electronic Engineering Department at the University of Bristol, Bristol, U.K. He leads the signal processing activities within the Centre for Communications Research of which he is Deputy Director. He has worked widely in the fields of 1-D and 2-D signal processing and has published over 250 papers, articles, and books. His current research is focused on the problems of image and video communications for both low bit rate wireless, Internet, and broadcast applications. He is a member of the EPSRC Communications College and the Steering Group for the DTI/EPSRC LINK program in Broadcast Technology.

Prof. Bull has served on the U.K. Foresight ITEC Panel and is a past Director of the VCE in digital broadcasting and multimedia technology. He is currently Chairman of a University spin-out company, ProVision Communication Technologies, Ltd., specializing in wireless multimedia communications. He is also currently involved in establishing a new DTI funded (£7.62 M) Centre for Communications Computing and Content that is based at Bristol. Prof. Bull is a Fellow of the IEE and a chartered engineer.