Automated Analysis of the Framing of Faces in a Large Video Archive

Graeme Phillipson
graeme.phillipson@bbc.co.uk

Ronan Forman
ronan.forman@bbc.co.uk

Mark Woosey
mark.woosey@bbc.co.uk

Craig Wright
craig.wright@bbc.co.uk

Michael Evans
michael.evans@bbc.co.uk

Stephen Jolly
stephen.jolly@bbc.co.uk

BBC Research & Development

Abstract

Automated editing systems require an understanding of how subjects are typically framed, and how framing in one shot relates to another. In this paper we present an automated analysis of the framing of faces within a large video archive. These results demonstrate that the *rule of thirds* alone is insufficient to describe framing that is typical in drama, and we show that the framing of one shot has an effect on that of the next.

1 Introduction

Automated editing systems [LC12][GRG14][MBC14][GRLC15][LDTA17] could enable broadcasters to provide coverage of more live events (such as music and arts festivals) where the cost of additional outside broadcast units would be prohibitive [WAC+18]. Constructing such systems requires an understanding of how to frame and sequence video. To frame video, systems often apply the *rule of thirds*, aligning faces on the dividing lines between the vertical and horizontal thirds [LC12][ST11]. More sophisticated approaches have been used, but these require large amounts of manually annotated data [SC14]. There is empirical evidence for the validity of the *rule of thirds*. However, this evidence also suggests that the rule does not fully explain how faces are framed [Cut15][WGLC17]. Additionally, it does not describe how framing in one shot relates to the next. In this paper we present an initial automated analysis of a large quantity of archive data, in contrast to previous investigations relying on human annotation. Manually-annotated data is assumed to be of a higher quality, and offers greater flexibility in what can be annotated. However, automated annotation is scalable to larger quantities of data, that may allow for more precise quantitative measures.

2 Data Gathering

The positions of faces in shots were found by analysing 8353 shows with a total duration of 8273 hours from a BBC archive. The genre of drama was chosen, on the assumption that the makers of these shows have more freedom to frame shots in an aesthetic manner. The shows were all broadcast in the UK between 2007 and

In: H. Wu, M. Si, A. Jhala (eds.): Proceedings of the Joint Workshop on Intelligent Narrative Technologies and Workshop on Intelligent Cinematography and Editing, Edmonton, Canada, 11-2018, published at http://ceur-ws.org

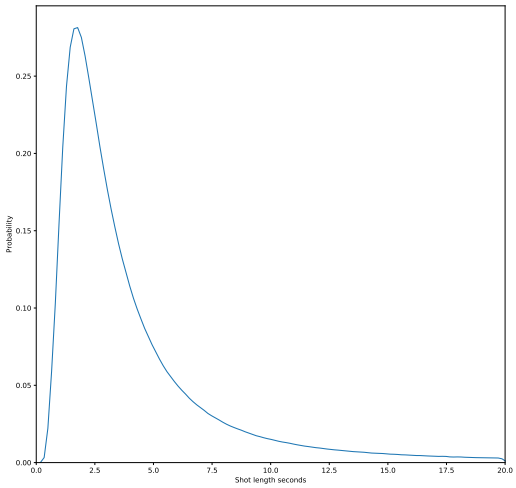


Figure 1: The probability of different shot lengths occuring

2018 in 16: 9 aspect ratio. Each was conformed to a resolution of 1024×576 before analysis. The first and last 5 minutes were trimmed from each show to remove trailers and title/credit sequences that may contain faces. Those faces might otherwise be found many times in the dataset and bias the results. The videos were split into discrete shots with *ffmpeg*[1]. The middle frame of each shot was extracted and assumed to be representative of the shot as a whole. We have not considered developing or action shots in this analysis, and they must be assumed to be adding some noise to the overall results. Shots shorter than 0.5s and longer than 20s were filtered out as they are likely to be the result of either false positive or negative shot change detections, or shots framed with subjects other than static faces in mind. The locations of the faces and their landmarks (e.g. the eyes) were found using the SeetaFace library [LKW+16]. Seetaface was chose because it's accuracy had been validated on this archive[IRF], which is important as not all off the shelf computer vision techniques generalise well enough to work across such a large archive. It is worth noting that SeetaFace will not detect partial faces, so we would not expect detections towards the very edge of the screen, where part of the face may be outside the visible frame. The centre of the face was taken to be the mid point between the eyes. $3,567,433$ faces were found in total.

3 Results

3.1 Shot distribution

The probability of different shot lengths can be seen in Fig.1. The mean shot length was 3.975s. The distribution shows a preference for shorter shots in most of the archive.

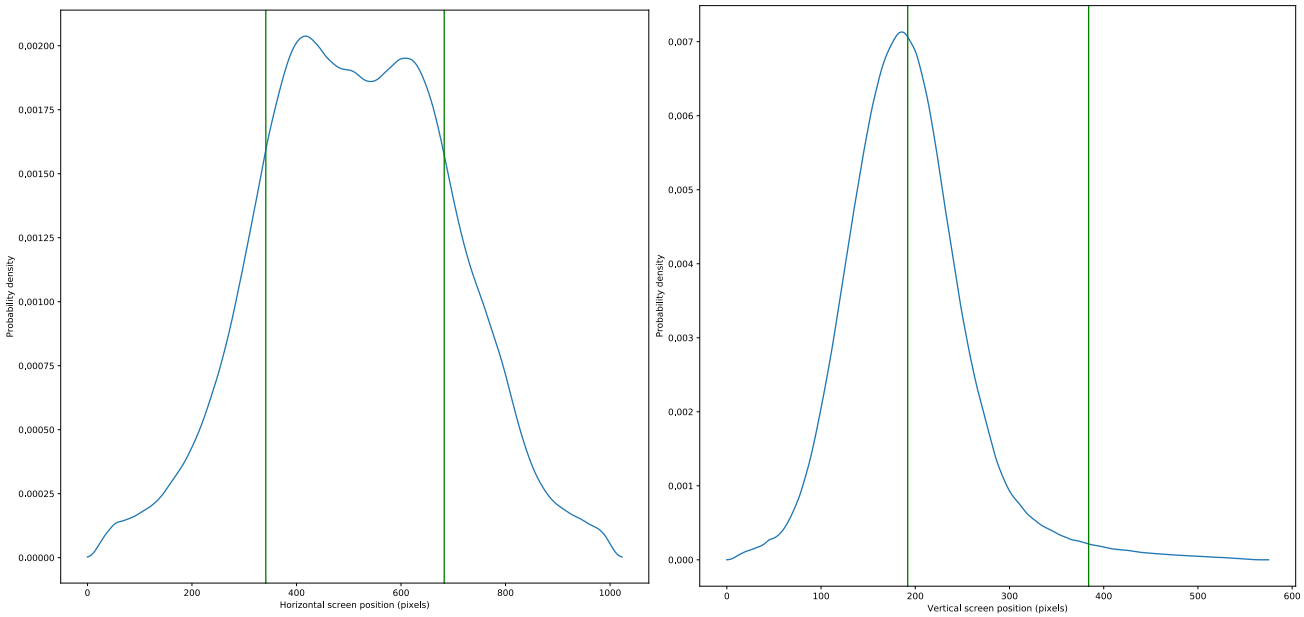
3.2 Head Position In All Shots

In Fig.2, the probability distribution of faces occurring at different locations within a shot is estimated across all the shots using Kernel Density Estimation [Sco15]. The vertical distribution shows a clear preference for the face to occur on the upper third line. The horizontal distribution shows a preference for faces to be within the middle third, particularly just inside the thirds lines, with a small preference for being on the left.

3.3 Head Position for Shots with Different Numbers of People

In Fig.3a the frequency of occurrence of faces in shots containing only one person is shown. There is a preference for the middle upper third line with two clusters at either end of this. There is also an asymmetric cluster to the right of and below the main cluster. Manual inspection of the shots responsible for this cluster shows that it is due to the presence of a overlaid sign language interpreter in a proportion of these shows, and whose face is

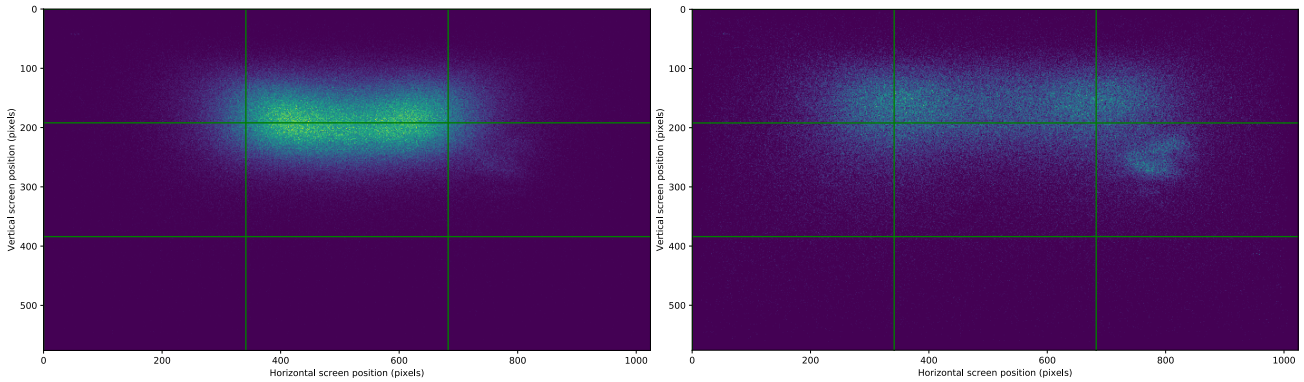
[1] https://www.ffmpeg.org/



(a) The density of faces across the horizontal axis.

(b) The density of faces across the vertical axis

Figure 2: The probability density of faces in shots across both the horizontal and vertical axes. The green lines represent the division into thirds.



(a) Singles

(b) Two-shots

Figure 3: The frequency of occurrence of face detection in each position within the analysed frames for shots with two people in them. The green lines indicate the *rule of thirds*

in approximately the same location in all of them. Fig.3b shows the same distribution for shots with two people in them. Here the average framing is slightly higher, and the two main clusters are spaced further apart.

3.4 Relationships Between Consecutive Shots

The relative framing of faces in two consecutive shots (where both shots contain only a single face) is illustrated in Fig.4. Given a face in a particular horizontal position (the x-axis) on the upper third line, the distribution of horizontal positions of faces in subsequent shots is as shown on the y-axis. For example, given a face located at 400px horizontally in one shot, the most likely position for the face in the subsequent shot is around 600px. This was calculated by storing all of the face detection locations in a KD-Tree [MM99], then walking a point across the upper third line. For each location on the line, all face detections within 10px were retrieved and the index of shots was used to find the locations of faces in the next shot. Kernel density estimation was then used to produce the conditional probability distribution of horizontal location in the next shot given the current horizontal position. The results show that when a face is in the left cluster it is likely that it will subsequently

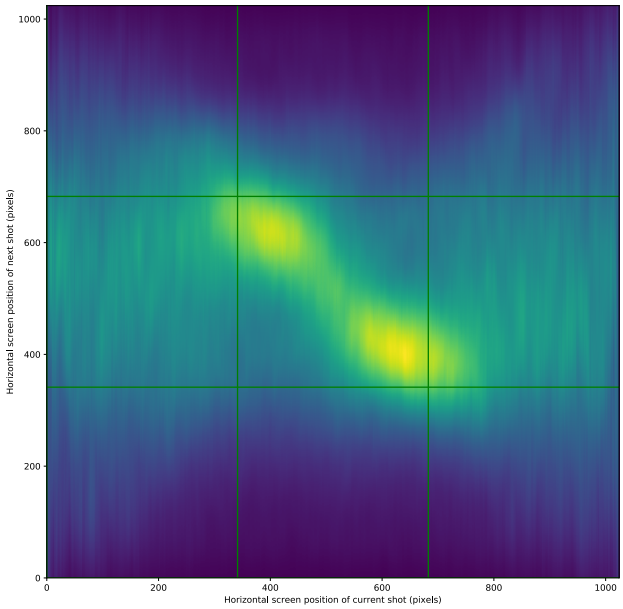


Figure 4: The conditional probability for the horizontal location of a face in the next (the y-axis) shot given a face in the current shot at a certain horizontal location, shown on the x-axis.

appear in the right cluster and vice versa.

3.5 Distribution of Face Sizes

The distribution of face sizes was calculated by taking the face landmarks produced by SeetaFace (the eyes, nose, and two corners of the mouth) and finding the convex hull of these points.[BDH96] The area of this convex hull was calculated and the distribution of this can be seen in Fig.5. Most production use a semi-standardised language to describe shots as being a "Close-up", "Mid-shot", "Long shot", etc. [ST11] The shots are defined in terms of where on the body the bottom of the screen cuts. If face area was strongly correlated with shot type then there might be a multi-modal distribution which could be used to estimate shot-type. However in Fig.5 we can see that while it is clearly not a single distribution, the overlap is too much to allow for shot type

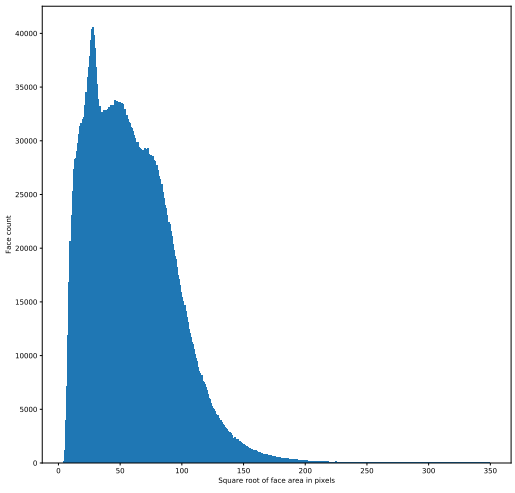


Figure 5: A histogram of the square root of the area of the face, as defined by the convex hull of the face landmarks.

classification.

4 Discussion

The results show that while the *rule of thirds* is important, there are deviations from it (such as the most likely face locations being slightly inside the lines for single shots, but on those lines for two-shots) which require large datasets in order to quantify.

Previous work has shown that single shots have a single centrally-framed cluster [Cut15][WGLC17] rather than the bimodal distribution demonstrated here. The bimodal distribution combined with the oscillations shown for the conditional probability of framing in consecutive shots suggests extensive use of the shot/reverse-shot pattern often used in dialogue [ST11]. The previous work concentrated on film, where as here we are examining television drama, and this difference in result may simply reflect how often the shot-reverse-shot pattern is used in these different media.

Expanding this work to analyse subjects other than faces is difficult, due to the lack of labelled data for this dataset to validate models other than simple face location. Particularly it is important to validate models on labeled data from this archive, as many open source systems were not trained on broadcast media. However, mass data labelling services [PBSA17] may provide a way to produce enough labeled data to validate other methods. This would allow visual features such as the framing of the whole body [RAG18][CSWS17][SJMS17][WRKS16] or salient non-human objects [CBSC18] to be investigated. Pose estimation would allow for investigation of the relationship between framing and the direction the direction of gaze. Dense pose estimation [RAG18] might be particularly useful as shot type are normally discussed in terms of where on the body the bottom of the frame cuts on the body which we would be able to calculate from this. Additionally this would allow the detection of people not facing the camera and, in turn, enable the detection of over the shoulder shots.

References

[BDH96] C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM TRANSACTIONS ON MATHEMATICAL SOFTWARE*, 22(4):469–483, 1996.

[CBSC18] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, 2018.

[CSWS17] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[Cut15] James E. Cutting. The framing of characters in popular movies. *Art & Perception*, 3(2):191–212, 2015.

[GRG14] Vineet Gandhi, Rémi Ronfard, and Michael Gleicher. Multi-Clip Video Editing from a Single Viewpoint. In *CVMP 2014 - European Conference on Visual Media Production*, page Article No. 9, London, United Kingdom, November 2014. ACM.

[GRLC15] Quentin Galvane, Rémi Ronfard, Christophe Lino, and Marc Christie. Continuity Editing for 3D Animation. In *AAAI Conference on Artificial Intelligence*, pages 753–761, Austin, Texas, United States, January 2015. AAAI Press.

[IRF] Irfs weeknotes 243. `https://www.bbc.co.uk/rd/blog/2017-05-irfs-weeknotes-number-243`. Accessed: 2018-10-3.

[LC12] Christophe Lino and Marc Christie. Efficient composition for virtual camera control. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '12, pages 65–70, Goslar Germany, Germany, 2012. Eurographics Association.

[LDTA17] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. Computational video editing for dialogue-driven scenes. *ACM Trans. Graph.*, 36(4):130:1–130:14, July 2017.

[LKW+16] Xin Liu, Meina Kan, Wanglong Wu, Shiguang Shan, and Xilin Chen. VIPLFaceNet: An open source deep face recognition sdk. *Frontiers of Computer Science*, 2016.

[MBC14] Billal Merabti, Kadi Bouatouch, and Marc Christie. A virtual director inspired by real directors. 2014.

[MM99] Songrit Maneewongvatana and David M. Mount. Analysis of approximate nearest neighbor searching with clustered point sets. *CoRR*, cs.CG/9901013, 1999.

[PBSA17] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153 – 163, 2017.

[RAG18] Iasonas Kokkinos Riza Alp Güler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. *arXiv*, 2018.

[SC14] Cunka Sanokho and Marc Christie. On-screen visual balance inspired by real movies. 2014.

[Sco15] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization, Second Edition*. Wiley, 2015.

[SJMS17] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[ST11] Roger Singelton-Turner. *Cue & Cut*. Manchester University Press, 2011.

[WAC+18] Craig Wright, Jack Allnut, Rosie Campbell, Michael Evans, Stephen Jollyand Lianne Kerlin, James Gibson, Graeme Phillipson, and Matthew Shotton. Ai in production: Video analysis and machine learning for expanded live events coverage. *Proceedings of the International Broadcasting Convention*, Sept 2018.

[WGLC17] Hui-Yin Wu, Quentin Galvane, Christophe Lino, and Marc Christie. Analyzing elements of style in annotated film clips. In *WICED 2017 - Eurographics Workshop on Intelligent Cinematography and Editing*, pages 29–35, Lyon, France, April 2017. The Eurographics Association.

[WRKS16] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.