High-level cinematic knowledge to predict inter-observer visual congruency

Alexandre Bruckert, Marc Christie

High-level cinematic knowledge to predict inter-observer visual congruency

Alexandre Bruckert
alexandre.bruckert@univ-nantes.fr
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004
F44000 Nantes, France

Marc Christie
marc.christie@irisa.fr
Inria, Univ Rennes, CNRS, IRISA
Rennes, France

ABSTRACT

When watching the same visual stimulus, humans can exhibit a wide range of gaze behaviors. These variations can be caused by bottom-up factors (i.e. features of the stimulus itself) or top-down factors (i.e. characteristics of the observers). Inter-observer visual congruency is a measure of this range. Moreover, it has been shown that cinematic techniques, such as camera motion or shot editing, have a significant impact on this measure [17]. In this work, we first propose a metric for measuring IOC in videos, taking into account the dynamic nature of the stimuli. Then, we propose a model for predicting inter-observer visual congruency in the context of feature films, by using high-level cinematic annotation as prior information in a deep learning framework.

CCS CONCEPTS

• **Computing methodologies** → *Interest point and salient region detections.*

KEYWORDS

gaze congruency, cinematography, neural networks

ACM Reference Format:
Alexandre Bruckert and Marc Christie. 2023. High-level cinematic knowledge to predict inter-observer visual congruency. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Predicting human viewing behavior, in the sense of gaze patterns or visual saliency for instance, is an important topic in the computer vision community. However, visual behavior is not always consistent between observers, either because of top-down factors (for instance, observers with a previous knowledge of the stimuli will exhibit different gaze patterns [5]), or bottom-up characteristics. For example, people will tend to exhibit very similar behaviors when viewing a scene containing a single salient object, while cluttered scenes, or scenes lacking strong visual attractors will induce more diversity in eye fixation locations. Differences in ages, or cultural background can also be found in visual attention data [4, 10]. Thus, understanding how different observers will react to a given stimulus is key in understanding the way we interact visually with images.

The similarity, or dissimilarity between visual trajectories among observers is referred as *attentional synchrony*, and metrics quantifying this synchrony are commonly called *inter-observer congruency* (IOC) metrics. Such metrics have proven very useful in a whole variety of applications, such as image ranking, quality assessment, or even visual saliency: indeed, IOC has been shown to provide an upper-bound on the performances of models predicting the locations of eye fixations. However, this measure in itself has received way less attention than, for instance, visual saliency prediction. Le Meur *et. al.* [9] offered a first image-processing approach, where they studied the influence of several image features, such as the depth of field or the image complexity, on IOC scores. Following this work, Rahman and Bruce [14] explored more image characteristics, coupled with top-down features. They proposed a predictive model of IOC based on both those feature sets, as well as information yielded by the predictions of visual saliency models. Bruckert *et. al.* [3] proposed an approach based on deep learning, relying on deep convolution networks to extract features, coupled with a shallow regression network. More recently, Yue *et. al.* [22] proposed a model for predicting IOC in the context of video, relying on a two-stream deep learning architecture.

In the context of movies, attentional synchrony has also been studied, from a more cognitive point of view. Dorr *et. al.* pointed out several differences in the variation of eye fixations and saccade amplitudes when watching the same stimulus several times over two days, and compared the synchrony observed on Hollywood movies and natural scenes. Mital *et. al.* [11] showed that the most predictive features for gaze clustering when viewing dynamic stimuli were temporal and motion-related, like flicker or contrast in motion. Smith and Mital [17] also studied the influence of the viewing task on attentional synchrony, highlighting a significant influence of it, but mostly after the first few fixations, which were usually guided by the exogenous attention mechanisms.

In the following, we first describe how to compute a reliable and suitable IOC score, inspired by visual saliency metrics, for dynamic stimuli. We then propose a model dedicated to predict this score in the context of feature films, by incorporating high-level priors using cinematic annotations.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 MEASURING INTER-OBSERVER CONGRUENCY

2.1 Previous metrics

A lot of methods have been proposed to describe the amount of visual congruency among observers when viewing a stimulus. All of those methods use different hypotheses about the distribution of gaze patterns, but overall, these metrics are highly correlated to one another [5]. Rajashekar *et. al.* [15] used the average z-score between the individual human fixations and the overall fixation density, using Kullback-Lieber divergence as a metric. Peters *et. al.* [12] used the normalized scanpath saliency metric (NSS) to compare each individual gaze track to a global inter-observer model, composed of the aggregation of individual saliency heatmaps.

Sawahata *et. al.* [16] used a criterion based on information theory, the entropy of the fixation distribution, or more precisely, the entropy of a Gaussian mixture model (GMM) fitted on the the gaze points divided into clusters based on the Bayesian information criterion (BIC). Similarly, Mital *et. al.* [11] used GMMs, and more specifically the weighted covariance value, to discriminate between "tightly and loosely clustered frames", i.e. frames in which attentional synchrony is higher or lower. Smith and Mital [17] also used these GMM clusters and their covariance, expressed as the visual angle enclosing 68% of gaze points. Finally, several area-based methods have been proposed: for instance, Goldstein *et. al.* [6] computed the area of the best-fit bivariate contour ellipse, whereas Breeden and Hanrahan [1] used the area of the convex hull of the fixation points.

Finally, more saliency-inspired methods consist in comparing the gaze tracks of a single observer to the joint distribution of all the other observers. This leave-one-out approach was used by Torralba *et. al.* [18] and Le Meur *et. al.* [9], where they use the rate of fixations falling in a saliency classifier, created from a thresholded fixation distribution map, and Rahman and Bruce [13], where they compute the AUC score between the individuals and the aggregated fixation distribution of all other observers.

2.2 Dynamic stimuli

Extending the measure of IOC to the spatio-temporal domain is not as straight forward as it may seem. For instance, applying an IOC measure on a frame-by-frame basis can be problematic, as there might not be enough fixations to avoid a significant amount of noise: indeed, in the case of cinematographic movies, each frame will be displayed for around 42 milliseconds, while the average eye fixation spans around a few hundred milliseconds, implying that each frame will only display one or two fixations per observer.

More generally, designing an IOC measure for dynamic stimuli implies answering questions about what we actually want to measure. For example, let us consider a sequence containing two spatially separate salient locations A and B (a dialogue between two characters, for instance), and two observers. If, during a short time period, the first observer fixates location A first and location B second, and the second observer does the opposite, both observers will exhibit similar spatial gaze patterns, and only differ temporally. However, a frame-by-frame measure will (in the worst scenario) treat the case as if the first observer only fixated location A and the second only location B. We then argue that a well-designed IOC metric should take into account the temporal continuity: two non-simultaneous fixations at the same spatial location should be considered as "close" based on the temporal dimension.

In order to address this issue, we propose a new approach to compute an IOC measure in the spatio-temporal domain.

First, we define the spatio-temporal fixation density map for a stimulus. For each frame, we compute the traditional fixation density map by convolving the binary fixation map with a Gaussian kernel, which covariance is chosen so that it approximates the size of the fovea. Figure 1 shows an example of this spatio-temporal representation. Then, we stack those density map into a spatio-temporal volume, and smooth it in the temporal dimension using a Gaussian kernel, which variance is set to approximate 250 ms, i.e. the average duration of a fixation. In the case of a 24 frames per second cinematic stimuli, this amounts to 6 frames. Now, this spatio-temporal map can be compared to ground truth fixations using the NSS metric on the whole volume:

$$\begin{aligned} NSS(S,F) &= \frac{1}{N}\sum_i \bar{S}_i F_i \\ \text{where} \quad \frac{1}{N} &= \sum_i F_i \quad \text{and} \quad \bar{S} = \frac{S-\mu(\hat{S})}{\sigma(S)} \end{aligned} \tag{1}$$

where N is the number of fixated voxels, S is the fixation density volume, F is a spatio-temporal binary fixation map, i.e. a volume where each voxel is either 1 if a fixation occurred at its location and time, and 0 otherwise. The choice of the NSS metric in this case comes straight forward, as it is way less time- and memory-consuming than AUC metrics.

Figure 1: Example of spatio-temporal fixation density map on a sequence of *Big Fish* (Tim Burton, 2004).

From there, we use the exact same leave-one-out approach than the static case. A fixation density is computed for each group of $(N_o - 1)$ observers, and compared using the NSS metric to the fixations of the remaining observer. The scores are then averaged over the observers to get a global IOC value. In order to track the evolution of attentional synchrony over time, we keep the global fixation densities and fixation maps, and compute the NSS values over a sliding time-window, which size can be chosen depending on the context: a shorter time window (e.g. four or five frames) allows for a finer-grained analysis, but is more sensitive to noise, for instance.

However, the main drawback of this method is its memory consumption. Indeed, we need to store a volume of size $H \times W \times T$ (where H is the height of the frame, W the width and T the duration of the whole sequence) for each observer, which can quickly

become overwhelming when working with high-resolution stimuli and (relatively) long movie sequences. In order to solve this issue, we designed a simple, yet useful heuristic.

We only consider a sliding time-window of size t; for each group of (N_o-1) observers, we gather all their fixations during this period, and report it on a 2D binary fixation map, which is then smoothed into a fixation density. This map is then compared to the binary fixation map of the remaining observer using the NSS metric. The process is iterated and averaged over all the observers to get an IOC score over the considered time frame. The duration of the time window can be freely chosen, once again depending on the context. In our analyses, we considered two window sizes: 5 frames for a fine-grain approach and 20 frames for a more general view. In our previous dataset [2], we found a strong significant correlation between this heuristic and the memory expensive approach (for time windows of 5 frames: $r = 0.7912$, $p < 0.001$; for time windows of 20 frames: $r = 0.8531$, $p < 0.001$). From now on, we will then only refer to this heuristic when we mention spatio-temporal IOC.

3 PREDICTING IOC FOR CINEMATIC STIMULI

In this section, we propose a bottom-up model dedicated to predict inter-observer visual congruency on dynamic stimuli, and more specifically on cinematic stimuli. For this purpose, we designed a two-stream deep neural network, inspired by the design of the ViNet saliency model [7].

3.1 Architecture

To design this model, we make the assumption that the features that drive attention in videos and that are extracted in deep saliency models should also play an important role into determining whether or not a stimulus will induce high or low visual congruency. This assumption was also made by Rahman and Bruce [14], with their Histogram of Predicted Salience features, where they use a stack of feature vectors extracted from several visual saliency models.

Our model is divided into three parts: (i) first, a two-stream encoder extracts features from the optical flow and the frames at different depths; (ii) then, similarly to the ViNet model [7], these features are passed through 3D convolution layers and upsampling, mixing the different hierarchical features using skip connections; (iii) finally, the resulting representation, alongside with IOC priors based on the cinematographic characteristics, is passed through fully connected layers to obtain an IOC value. The overall architecture is shown on Figure 2.

Two-stream encoder : The encoder part is composed of two S3D networks [21], one for the spatial features, using a stack of 32 consecutive frames as the input, and the other using the same 32 stack with optical flow. Following the approach of ViNet [7], for a frame at time t, the input is composed by the frames $F_{t-32+1}, ..., F_t$ and the optical flow maps $O_{t-32+1}, ..., O_t$.

The features are extracted at the end of the four convolution blocks, and passed through skip connections to the decoder module, at different hierarchical levels. For an input of shape $[T \times C \times H \times W]$, where T is the time window (in our case, 32), C is the number of channels of the input (in our case, 3) and H and W are the height and width of the considered frame, the four features vectors, X_1, X_2, X_3 and X_4 have respective shapes of $[192 \times 16 \times \frac{H}{4} \times \frac{W}{4}]$, $[480 \times 16 \times \frac{H}{8} \times \frac{W}{8}]$, $[832 \times 8 \times \frac{H}{16} \times \frac{W}{16}]$ and $[1024 \times 4 \times \frac{H}{32} \times \frac{W}{32}]$.

Decoder module : The decoder module consists in a succession of concatenations alongside the temporal axis, gathering the hierarchical features from the two stream and the output of the previous upsampling layer, 3D convolution layers, and upsampling using trinlinear interpolation. This integration part is then followed by three 3D convolution layers, to reduce the feature tensor to one in the channel and temporal dimensions. The output features are then flattened, batch-normalized and concatenated with IOC priors, before being passed through three dense layers (similarly to the static IOC model) of size 1024, 256 and 1.

Cinematic IOC priors : In a previous work [2], we showed that high-level cinematic information can significantly influence inter-observer visual congruency, and is most likely not taken into account by the feature extractor. We then include five prior values into the feature vector:

- A camera motion prior, which is the average IOC value for the type of camera movement in the shot of the considered frame,
- A shot size prior, which is the average IOC value for the shot size of the considered frame,
- A shot angle prior, which is the average IOC value for the shot angle of the considered frame,
- The entropy of the flicker map of the considered frame,
- A cut prior, which is the average IOC value of frames within the first 500 milliseconds following a cut if the frame is in this situation, and the average IOC value of the other frames if not.

In their work, Mital *et. al* [11] showed that flicker, i.e. the change in luminance over time, alongside with motion, is also a strong predictor of gaze clustering. Since motion is already taken into account by the optical flow stream, we include flicker by computing the entropy of a flicker map: at time t, we consider frames $F_{t-4}, ..., F_t$, and transfer them from RGB to the CIELAB color space. We then compute the absolute difference of the frames luminance values $(L_{t-4}, ..., L_t)$, and average it:

$$Fl_t = \frac{1}{N} \sum_{i=1}^{N} |L_{t-i} - L_{t-i+1}| \tag{2}$$

Where Fl_t is the flicker map at time t and N is the number of successive frames considered. In our case, we use $N = 5$, similarly to Smith and Mital [17], in order to minimize the influence of noise due to compression artifacts.

3.2 Training

3.2.1 Implementation details. The frames are first resized to $[288 \times 512]$, using letterboxing if needed to respect the original aspect ratio of the frame. The optical flow frames are processed using the same procedure as Xie *et. al.* [21]: the optical flow is extracted using the TV-L1 algorithm [23], the magnitude is truncated into $[-20, 20]$, and the maps are then stored as 3-channels encoded JPEG files.

To process the frame F_t, the sequence $F_{t-32+1}, ..., F_t$ is fed to the model. If any of those frames fall before the first frame of the clip, the first frame is just repeated the adequate amount of times. In

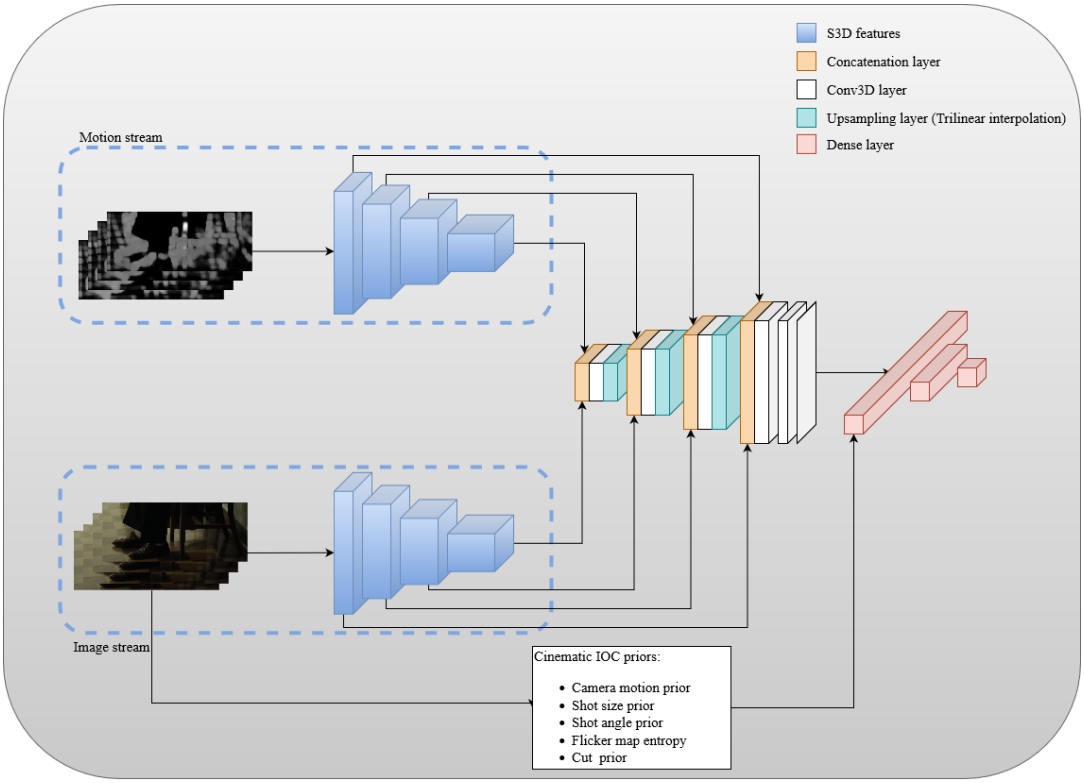

Figure 2: Architecture of the proposed dynamic IOC model

order to train the network, we select the 32-frames sequences in a random order among all clips

The priors are computed based on available information; if no editing annotation is provided, we take the average IOC value of the whole dataset for each IOC prior.

The S3D encoder are initialized using weights pre-trained on the Kinetics dataset [8] on an action-recognition task, using both RGB frames and optical flow. We use the L2 norm as a loss function, with the Adam optimizer, learning rate is initially set at $10e-4$, and the batch size is set at 4.

3.2.2 Training datasets. The model is first trained on the DHF1k dastaset [19]. Ground truth IOC scores are computed based on the supplied scanpaths (using the 20-frames time window). The 500 first clips from the training set are used for training, and the remaining 100 are used for validation, and for early stopping. While the Hollywood2 dataset would have been useful to train on, as it features the type of clips we are interested in, its limitations prevented us from using it. The low number of free-viewing observers makes it difficult to get a reliable IOC score, and, while adding task-oriented data can be useful for visual saliency, it induces too much of a bias for IOC prediction.

Then, we use 15 clips from our dataset to fine-tune the model (12 for training, 3 for validation), using the IOC priors as we have cinematographic annotations, holding out the 5 remaining clips for testing purposes.

3.3 Results

We used three datasets to evaluate the model: the validation set of DHF1k (100 clips), the 5 held out clips from our dataset, and the dataset from Breeden and Hanrahan [1].

We observe a Pearson correlation coefficient score between the predicted IOC values and the ground-truth of $r = 0.691$ ($p < 10^{-5}$) for the DHF1k dataset, $r = 0.731$ ($p < 10^{-5}$) for Breeden's dataset and $r = 0.755$ ($p < 10^{-5}$) for ours. These scores are much higher than those obtained with static models [3, 14], which can be explained by the prominent role played by motion features on IOC [11]. DHF1k results also seem to be lower than the other, probably due to the absence of cinematographic priors and annotations, that are used in Breeden's and our dataset.

3.3.1 Ablation study. In order to evaluate how each part of the model contributes to the overall performances, and especially how the cinematic priors have an influence, we performed and ablation study, retraining different settings of the model. First, we tried both branches (RGB and Optical Flow) separated, without any priors. Then, we use the two streams and all of the priors but one each time: the camera motion prior (1), the shot size prior (2), the shot angle prior (3), the flicker map entropy (4) and the cut prior (5). Results for each configuration is shown in Table 1

As expected, on the DHF1k set, as there is no significant prior, the correlation scores do not vary when removing priors, except in configuration (4), where the entropy of the flicker map is removed. The camera angle prior does not seem to have any impact on the

Dataset	DHF1k [20]	Breeden [1]	Ours
RGB-stream (no prior)	0.631	0.624	0.657
Flow-stream (no prior)	0.471	0.473	0.469
Two-stream+priors (1)	0.690	0.712	0.733
Two-stream+priors (2)	0.689	**0.731**	0.728
Two-stream+priors (3)	0.690	**0.731**	0.754
Two-stream+priors (4)	0.652	0.699	0.718
Two-stream+priors (5)	**0.691**	0.707	0.743
Full model	**0.691**	**0.731**	**0.755**

Table 1: Pearson correlation coefficient between predicted IOC scores and ground truth IOC for several models

prediction, which is consistent with what we observed in previous work [2], and can probably be removed. A small improvement is seen when adding the optical flow stream to the RGB stream. The relatively low value for this improvement can be explained by the fact that the RGB-stream already extract at least some motion features, because of its 3D-CNN feature extractor. Finally, overall, adding cinematographic high-level information through these priors seems to be of interest for predicting inter-observer visual congruency.

4 DISCUSSION

In this work, we focused our attention on inter-observer visual congruency, a measure of how similar gaze behaviors from different observers are when they are watching the same stimulus. We proposed a way to measure this phenomenon on dynamic stimuli, and introduced a model to predict it on movie sequences.

While inter-observer congruency (or attentional synchrony) is well known and studied by cognitive psychologists, we argue that more attention should be payed to this measure in computer vision, both from a modeling point of view and for the resulting applications. While its role as an upper bound of the performance of visual attention models is well-known, it can also be used to constraint visual saliency predictions: for instance, a predicted saliency map exhibiting a lot of salient areas will probably be wrong if the IOC is high (meaning that observers tend to look at the same place). In this regard, predicting IOC can be used to give an estimation of how "difficult" a saliency prediction will be, and serve as a likelihood score.

It could also be interesting to evaluate the interest of this measure in the context of image quality assessment: a high degree of visual congruency means that there might be a single strong visual attractor on the image, and thus artifacts on other areas of the frame could be overlooked.

From the perspective of filmmaking, knowing when viewers will focus their attention in the same location is tremendously useful for directors, as it allows them even more control on what the viewer experiences, in order to convey their narrative content and messages at best. For virtual cinematography and automated editing, this can be used to constraint the choice of the cuts, for instance, depending on the desired style.

REFERENCES

[1] Katherine Breeden and Pat Hanrahan. 2017. Gaze Data for the Analysis of Attention in Feature Films. *ACM Transactions on Applied Perception* 14, 4 (2017). https://doi.org/10.1145/3127588

[2] Alexandre Bruckert, Marc Christie, and Olivier Le Meur. 2022. Where to look at the movies: Analyzing visual attention to understand movie editing. *Behavior Research Methods* (2022), 1–20.

[3] Alexandre Bruckert, Yat Hong Lam, Marc Christie, and Olivier Le Meur. 2019. Deep Learning For Inter-Observer Congruency Prediction. In *2019 IEEE International Conference on Image Processing (ICIP)*. 3766–3770. https://doi.org/10.1109/ICIP.2019.8803596

[4] Hannah F. Chua, Julie E. Boland, and Richard E. Nisbett. 2005. Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences* 102, 35 (2005), 12629–12633. https://doi.org/10.1073/pnas.0506162102

[5] Michael Dorr, Thomas Martinetz, Karl R. Gegenfurtner, and Erhardt Barth. 2010. Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision* 10, 10 (2010), 28–28.

[6] Robert B. Goldstein, Russell L. Woods, and Eli Peli. 2007. Where people look when watching movies: do all viewers look at the same place? *Computers in Biology and Medicine* 37, 7 (2007), 957—-964. https://doi.org/10.1016/j.compbiomed.2006.08.018

[7] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian, and Vineet Gandhi. 2021. ViNet: Pushing the limits of Visual Modality for Audio-Visual Saliency Prediction. arXiv:2012.06170 [cs.CV]

[8] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950 [cs.CV]

[9] Olivier Le Meur, Thierry Baccino, and Aline Roumy. 2011. Prediction of the Inter-Observer Visual Congruency (IOVC) and Application to Image Ranking. In *Proceedings of the 19th ACM International Conference on Multimedia*. 373–382. https://doi.org/10.1145/2072298.2072347

[10] Olivier Le Meur, Antoine Coutrot, Zhi Liu, Pia Rämä, Adrien Le Roch, and Andrea Helo. 2017. Visual Attention Saccadic Models Learn to Emulate Gaze Patterns From Childhood to Adulthood. *IEEE Transactions on Image Processing* 26, 10 (2017), 4777–4789. https://doi.org/10.1109/TIP.2017.2722238

[11] Parag K. Mital, Tim J. Smith, Robin L. Hill, and John M. Henderson. 2011. Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion. *Cognitive Computation* 3, 1 (2011), 5––24. https://doi.org/10.1007/s12559-010-9074-z

[12] Robert J Peters, Asha Iyer, Laurent Itti, and Christof Koch. 2005. Components of bottom-up gaze allocation in natural images. *Vision research* 45, 18 (2005), 2397–2416.

[13] Anis Rahman, Denis Pellerin, and Dominique Houzet. 2014. Influence of number, location and size of faces on gaze in video. *Journal of Eye Movement Research* 7, 2 (2014), 891—-901. https://doi.org/10.16910/jemr.7.2.5

[14] Shafin Rahman and Neil D. B. Bruce. 2016. Factors Underlying Inter-Observer Agreement in Gaze Patterns: Predictive Modelling and Analysis. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16)*. 155–162. https://doi.org/10.1145/2857491.2857495

[15] Umesh Rajashekar, Lawrence K. Cormack, and Alan C. Bovik. 2004. Point-of-gaze analysis reveals visual search strategies. In *Human Vision and Electronic Imaging IX*, Vol. 5292. SPIE, 296 – 306. https://doi.org/10.1117/12.537118

[16] Yasuhito Sawahata, Rajiv Khosla, Kazuteru Komine, Nobuyuki Hiruma, Takayuki Itou, Seiji Watanabe, Yuji Suzuki, Yumiko Hara, and Nobuo Issiki. 2008. Determining comprehension and quality of TV programs using eye-gaze tracking. *Pattern Recognition* 41, 5 (2008), 1610–1626. https://doi.org/10.1016/j.patcog.2007.10.010

[17] Tim J. Smith and Parag K. Mital. 2013. Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision* 13, 8 (2013), 16–16. https://doi.org/10.1167/13.8.16

[18] Antonio Torralba, Monica S. Castelhano, Aude Oliva, and John M. Henderson. 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review* 113 (2006), 766–786.

[19] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. [n. d.]. Revisiting Video Saliency Prediction in the Deep Learning Era. https://mmcheng.net/videosal/.

[20] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. 2019. Revisiting Video Saliency Prediction in the Deep Learning Era. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019). https://doi.org/10.1109/TPAMI.2019.2924417

[21] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In *ECCV 2018*. 318–335. https://doi.org/10.1007/978-3-030-01267-0_19

[22] Jiaomin Yue, Qiang Lu, Dandan Zhu, Xiongkuo Min, Xiao-Ping Zhang, and Guangtao Zhai. 2021. Inter-Observer Visual Congruency in Video-Viewing. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*. 1–5. https://doi.org/10.1109/VCIP53242.2021.9675428

[23] C. Zach, T. Pock, and H. Bischof. 2007. A Duality Based Approach for Realtime TV-L1 Optical Flow. In *Pattern Recognition*. 214–223. https://doi.org/978-3-540-74936-3