Deep Spatio-Temporal Modeling for Object-Level Gaze-Based Relevance Assessment

Konstantinos Stavridis, Athanasios Psaltis, Anastasios Dimou, Georgios Th. Papadopoulos and Petros Daras

Centre for Research and Technology Hellas

{staurid,at.psaltis,dimou,papad,daras}@iti.gr

Abstract—**The current work investigates the problem of object-level relevance assessment prediction, taking into account the user's captured gaze signal (behaviour) and following the Deep Learning (DL) paradigm. Human gaze, as a sub-conscious response, is influenced from several factors related to the human mental activity. Several studies have so far proposed methodologies based on the use of gaze statistical modeling and naive classifiers for assessing images or image patches as relevant or not to the user's interests. Nevertheless, the outstanding majority of literature approaches only relied so far on the use of hand-crafted features and relative simple classification schemes. On the contrary, the current work focuses on the use of DL schemes that will enable the modeling of complex patterns in the captured gaze signal and the subsequent derivation of corresponding discriminant features. Novel contributions of this study include: a) the introduction of a large-scale annotated gaze dataset, suitable for training DL models, b) a novel method for gaze modeling, capable of handling gaze sensor errors, and c) a DL based method, able to capture gaze patterns for assessing image objects as relevant or non-relevant, with respect to the user's preferences. Extensive experiments demonstrate the efficiency of the proposed method, taking also into consideration key factors related to the human gaze behaviour.**

Index Terms—**Gaze modeling, DL, relevance assessment**

I. Introduction

Implicit human feedback, both conscious and sub-conscious, can provide valuable information, among others, for customizing interfaces, improving user experience and optimizing computer vision algorithms. The gaze signal is one of the most informative sources of implicit sub-conscious human response. Gaze patterns inherently exhibit a high level of complexity, as they can be influenced by a large number of parameters/factors, such as the type of the observation task (free-viewing, target search, etc.), the user's mental state and the user's viewing patterns. Therefore, gaze modeling and eventual interpretation comprises a rather challenging problem that needs to be investigated under well defined experimental settings, including the actual definition of the experiment as well as the surrounding environmental conditions.

The current work addresses the challenge of classifying objects present in an image as relevant or non-relevant to the user's target search, by taking only into account and interpreting the captured user's gaze signal during a search session. Key methodological selection of the current work is the development of appropriate Deep Learning (DL) techniques [18] for modeling the spatio-temporal patterns of the captured users' gaze signal, due to their reported ability in efficiently handling complex pattern detection and classification tasks, while also demonstrating outstanding generalization capabilities. However, large-scale Neural Networks (NNs) typically require large-scale annotated datasets for training purposes. The main contributions of the current study are: a) **a large-scale annotated gaze dataset** that involved 48 participants during the capturing phase, whose scale is appropriate for supporting the development of DL methods, b) **a novel method for modeling gaze patterns** is proposed, which does not take into account only individual object-related information, but also considers surrounding contextual cues, and c) **a new DL-based relevance assessment classifier** is proposed that makes use of the proposed gaze modeling in order to assess the observed objects as relevant or not to the user's search intentions.

The rest of the paper is organized as follows: Section II describes the work related to image/object-level gaze-based relevance assessment. The introduced large-scale annotated gaze dataset is presented in Section III. Section IV details the proposed gaze modeling and the respective NN-based classifier. Experimental evaluation is discussed in Section V and conclusions are drawn in Section VI.

II. Related Work

Gaze has attracted the scientific interest of researchers [5], [16], [17], [19]. In the context of image relevance assessment prediction using gaze data, several studies have been presented so far. The authors of [11] propose a method that utilizes the gaze pattern for inferring relevance feedback to be used for image retrieval, according to the user's preferences, while using statistical metrics as features and logistic regression for the relevance prediction. The gaze signal is modeled at the fixation level by using features such as mean length of fixations, total length of fixations, standard deviation of fixation occurrence times, times of revisiting an image, etc. In [21], image relevance assessment is used for realizing image tagging. Gaze features, such as the the total number of fixations, the total length of fixations and the maximum pupil diameter are used for that purpose. The proposed method in [22], uses visual features combined with spatial gaze modeling (heatmaps) and adopts the DL paradigm for the

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

image relevance assessment prediction. It must be noted that relevance assessment, regarding the aforementioned studies, is performed at the image level and not at the object level, compared to the focus of the current study.

Fewer studies through the literature explore the image objects' relevance assessment. According to the study [10], few objects in a video sequence are augmented by information boxes in the video frames. Gaze measurements are recorded so as to infer the relevancy of objects. The logistic regression classifier is selected and the respective features are defined making use of statistical metrics, such as average and sum of fixations on objects and text information boxes. Another approach that operates at the object level is presented in [15]. The authors introduce means of handling temporal and spatial elements of gaze energy distribution. The temporal features take into account when fixations are located on the object area. The spatial features are computed by calculating the sum of the gaze energy in concentric rings, where the center is defined by the object center of gravity. The proposed features are then fed to a SVM classifier for predicting the relevance assessment. The study attempts to capture temporal patterns in the gaze signal, but in an aggregating and averaging way, i.e. it does not aim to capture gaze time evolving patterns. Taking into account the above-mentioned analysis, it can be seen that the wide majority of literature approaches only rely on the use of explicitly defined, simple and hand-crafted features, combined also with relatively simple or naive classification schemes. Additionally, most of the approaches, proposed so far, only constrain their analysis at the image level. Nevertheless, the recent advances in the machine learning community (and in particular the so called DL paradigm) for automatically learning optimal task-oriented features, capable of modeling and supporting the detection of complex and hierarchical patterns, have not yet been extensively investigated in the field of gaze-based relevance assessment prediction. Additionally, so far the potential of incorporating contextual information (and in particular the presence and particular spatial configuration of the objects in an image) has not been examined.

III. Introduced Gaze Dataset

Some annotated gaze datasets have been created so far, to study the saliency prediction problem [2], [4], [7]–[9], [20]. These have been created by adopting the free observation paradigm, where the human subjects were free to look wherever they wanted to, during the capturing session. On the other hand, the dataset proposed in [3] adapts to the memory task, according to which after the completion of the observation step the subjects were asked to mention what they had seen. The aforementioned gaze dataset creation methodologies do not require subjects to look at specific targets (target search paradigm), which is what it is actually required for the relevance assessment prediction task. In [19] a task-driven gaze annotated dataset, UPMC-G20, is introduced. Nevertheless, none of the mentioned datasets has adequate size to be used for training DL models. Taking into account the above, a large-scale annotated gaze dataset, specifically tailored to the visual target, search paradigm, is required and, to this end, constitutes one of the central goals of this study. The gaze annotated dataset is publicly available at: *https://vcl.iti.gr/dataset/gata/*.

In order to create an annotated gaze dataset, a simple interface (visual stimulus) that included 6 images (center-aligned in two rows), denoted as image-group session, was projected on a $23''$ monitor. A gaze sensor (myGaze) [14], having gaze position accuracy of 0.5 degrees and spatial resolution of 0.1 degrees, was placed at the bottom-center of the screen for capturing the observed subjects gaze behaviour, namely a gaze point trajectory of the user's look on the monitor plane. The sensors capturing frequency was 30Hz. Additionally, a wireless mouse sensor was handed to the user, in order to freely indicate the beginning and the end of the capturing image-group session. The utilized images constituted general-purpose ones and were randomly selected from the publicly available COCO dataset [12]. Before the beginning of each image-group session, the human subject was given a keyword (object type) as a query term and was subsequently asked to search for instances of this category of objects in the depicted images. The set of supported objects consisted of 80 types of every-day ones, such as persons, cars, cats, etc. Each subject was positioned approximately $\sim$ 60-70cm away from the computer monitor, as depicted in Fig. 1, and was informed about the capturing procedure prior to the execution of the experiments. For each human subject, the sensor was calibrated to the subjects eyes, before any gaze capturing taking place. Moreover, each human subject underwent subsequent capturing image-group sessions (i.e. different sets of 6 images included in the projected interface with different query terms defined each time), where the total duration of all sequential capturing image-group sessions did not exceed the limit of $15 - 20$ minutes (which corresponded to a targeted number of approximately $\sim$ 85 image-group sessions per experimental session). Overall, forty eight (48) individual subjects, 41 males and 7 females, were involved in the gaze recordings, aging from 22 to 45, while a total number of 238 experimental sessions were captured (i.e. 20.000 image-group sessions).

The assembled dataset comprises a large-scale benchmark of approximately 120.000 object instances with associated gaze signal captured, given a query object class. In general, the COCO dataset includes images with a varying number of objects with different sizes and occlusion levels. Therefore, the created dataset comprises a challenging one for interpreting the human gaze signal and also investigating its possible integration in image analysis methods. In terms of gaze characteristics, the minimum, maximum and average number of gaze points (captured at a frequency of 30Hz) per image are 1, 1.842 and 42.251, respectively. In terms of fixations, the corresponding numbers are 1, 104 and 3.166, respectively. The terms "gaze-point" and "fixation" are defined in detail in Section IV.

Fig. 1. Gaze capturing setting

IV. Proposed Method

A. Gaze modeling

Capturing gaze patterns is not a trivial task, since the gaze signal typically contains large amounts of noise, while it is also significantly affected by human (mental state) and environmental (distractions) factors. In this respect, a robust, detailed and efficient gaze modeling approach is required, which will also effectively capture both its spatial and temporal characteristics.

Encoding sequential information in the form of vector representations (embeddings) has been widely used in the field of text mining [13]. Following the same principles, gaze can be modeled as a series of N-dimensional vectors. Each vector will contain spatial information, with respect to a specific time instant. For encoding spatial information, the terms "gaze-point" and "fixation" are initially defined. In particular, gaze-point is the point on the screen where the human gaze has been recorded in one time unit. Fixation is defined as the circular image area with radius R, where subsequent gaze-points are concentrated for a minimum time interval T. In the current study, the fixation parameters, as explained in [15] and considering the relatively small number of fixations per image, were set as follows: $R = 30$ pixels and $T = 100$ msec.

Generally, the gaze sensor sensitivity/accuracy, the user's mental state and stress level, the inevitable rapid head movements, potential environmental factors (e.g. distractions) introduce noise to the captured gaze signal. The resulting erroneous displacements of the fixations on the monitor need to be handled by appropriate modeling techniques. On the other hand, the spatial configuration of objects in an image influences the observed gaze attention. The latter factor is termed 'object context' in this work and refers to the objects' surrounding environment. Fundamental consideration of the proposed gaze modeling approach is to include such object contextual information, in order to improve the object relevance assessment predictions. In particular, the potential of using the distance between the fixations and the objects is introduced for addressing the above requirements. The intuition behind this choice is that a relatively small distance of a displaced fixation from the stared object can indicate whether the fixation is associated with the object or not. The considered fixation-to-object distance is normalized by the object's size, since the latter influences visual attention. Taking into account the normalized distances of each fixation from all objects in an image, object context is included in the modeling. Under the proposed approach, the gaze signal can be treated as a sequence of fixations or gaze-points, depending on the adopted level of abstraction. In this respect, for each fixation, the normalized distances from all objects in the image are estimated as follows:

$$d_{F_iO_j} = \frac{D_{F_iO_j}}{S_j}, \tag{1}$$

where $d_{F_iO_j}$ is the normalized distance of fixation F_i from object O_j, $D_{F_iO_j}$ is the Euclidean distance of fixation F_i from the center of the bounding box of object O_j and S_j is the size of object O_j. Object size is defined as the Euclidean distance of its center from the bottom-right vertex of its bounding box, as also depicted in Fig. 2.

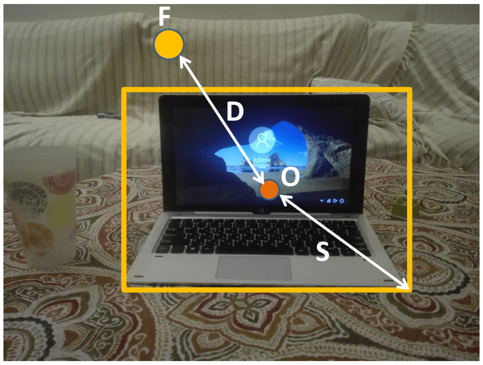

Fig. 2. Object Size - Fixation to Object Distance.

Taking into account the above consideration and definitions, the first proposed modeling approach, so called 'Fixation Duration - Distance Embedding (FD-DE)', operates at the fixation level and considers explicitly the duration of each observed F_i. The remaining elements of the proposed embedding comprise the normalized distances $d_{F_iO_j}$ of fixation F_i from each image object O_j. Overall, the proposed embedding FD-DE comprises a sequence of vectors (one for every fixation F_i), where each such vector includes the value of the duration of F_i (denoted T_{F_i}) and its normalized distances from all image objects ($d_{F_iO_j}$), as illustrated in Table I. It needs to be highlighted that in the FD-DE representation, duration is only indicated through the values of T_{F_i}, i.e. through a direct registration of each fixation duration. The proposed modeling approach can be also applied to the gaze-point level of abstraction, by excluding the duration element, ['Gaze-Point Distance Embedding' (GP-DE)].

TABLE I
'Fixation Duration - Distance Embedding (FD-DE)' gaze modeling approach

F_1	F_2	F_3	...	F_n
T_{F_1} $d_{F_1O_1}$ $d_{F_1O_2}$ ... $d_{F_1O_m}$	T_{F_2} $d_{F_2O_1}$ $d_{F_2O_2}$ ... $d_{F_2O_m}$	T_{F_3} $d_{F_3O_1}$ $d_{F_3O_2}$ ... $d_{F_3O_m}$	... ...	T_{F_n} $d_{F_nO_1}$ $d_{F_nO_2}$ ... $d_{F_nO_m}$

An alternative approach, so called 'Fixation Allocation - Distance Embedding (FA-DE)' that operates only at the fixation level is also proposed, which, however, emphasizes

TABLE II
'FIXATION ALLOCATION - DISTANCE EMBEDDING (FA-DE)' GAZE MODELING APPROACH

F_1	F_2		...	F_n	
G_1	G_2	G_3	...	G_{99}	G_{100}
$d_{F_1O_1}$ $d_{F_1O_2}$ ... $d_{F_1O_m}$	$d_{F_2O_1}$ $d_{F_2O_2}$ ... $d_{F_2O_m}$	$d_{F_2O_1}$ $d_{F_2O_2}$ ... $d_{F_2O_m}$	... ...	$d_{F_nO_1}$ $d_{F_nO_2}$ ... $d_{F_nO_m}$	$d_{F_nO_1}$ $d_{F_nO_2}$ ... $d_{F_nO_m}$

TABLE III
COMPARATIVE RESULTS

Gaze Modeling	Accuracy	Relevant Rate	Non-relevant Rate
FD-DE	0.7274	0.2443	0.8794
FA-DE	0.7357	0.1942	0.9062
GP-DE	**0.9476**	**0.5577**	**0.9584**

more on the temporal allocation of the observed fixations. In particular, the overall capturing session duration (when only time corresponding to fixations is considered) is divided into K equal length slots G_k ($K = 100$ based on experiments in this work). Then, for each slot G_k the respective active F_i is considered and the corresponding distance embedding (i.e. normalized Euclidean distances $d_{F_iO_j}$ from all image object O_j) are computed, similarly to the case of the FD-DE representation. Based on the aforementioned definition, the distance embeddings for subsequent G_k that correspond to the same F_i are essential vectors with same values; however, as it will be experimentally verified in Section V, this fact does not affect the discrimination capability of the FA-DE representation. The proposed FA-DE representation is illustrated in Table II.

B. Relevance assessment prediction

The goal of this study is to analyze the human gaze behaviour, in order to efficiently predict the relevance of the observed objects, compared to the search target/criteria of the human subject. According to the study in [1], humans recognize visual content by repeatedly focusing the fovea to subsequent relevant image areas (objects), while also aggregating the awareness of the identified entities (objects) in the ongoing visual analysis task. Building on this fundamental consideration, Recurrent Neural Networks (RNNs), which are ideal for modeling time evolving processes, are suitable for efficiently capturing and modeling the temporal characteristics and inter-dependencies of the gaze signal, i.e. encoding them in the NN internal and, in this way, simulating the fundamental functionalities of the human vision system.

Under the proposed approach, Long-Short-Term-Memory (LSTM) networks [6], i.e. a particular type of RNN with increased long-term temporal modeling capabilities, are used. The proposed NN architecture is based on a two-level LSTM network, so as to be capable of capturing complex hierarchical gaze data inter-dependencies. The proposed gaze embeddings FD-DE, FA-DE and GP-DE (Section IV) are used as input sequences to the proposed LSTM architecture. On top of the LSTM layers, a dense one is integrated, which receives as input the internal state vector of the last LSTM and its size is equal to the maximum number of objects in the image (as estimated in the training set). This dense layer eventually classifies all image objects as relevant or not, taking into account the user's behaviour and with respect to the target search. It must be noted that the ordering of the predicted objects (O_1,O_2,..,O_m) is identical to the ordering of fixation-to-object distances ($d_{F_iO_1}$, $d_{F_iO_2}$,..,$d_{F_iO_m}$). The latter enables the NN to encode the respective correlations. The overall proposed architecture is presented in Fig. 3.

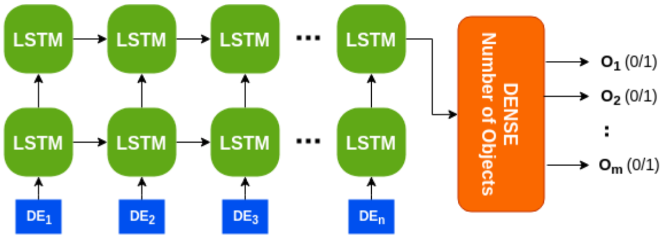

Fig. 3. Proposed DL architecture for user relevance assessment prediction.

V. EXPERIMENTAL RESULTS

Every image object, derived from COCO annotation data, is considered as relevant (assigned a value equal to '1') if it belongs to the search object class of the respective capturing session or as non-relevant (value equal to '0') otherwise. In order to handle the varying number of objects per image, the total number of objects that the proposed model processed is set equal to the maximum number of objects found in an image of the utilized dataset. Objects that do not belong to the dataset supported classes are neglected during the evaluation, while their relevance degree is set to '0' and all involved distances to '-1'.

With respect to the gaze modeling approaches, namely the FD-DE, FA-DE and GP-DE, Table III provides quantitative object detection results. In particular, the FD-DE modeling performs better than FA-DE in terms of relevant rate. The initial hypothesis was that extending the sequence of fixation embeddings, by repetition based on duration ratio, would improve the performance, however from the presented results it can be observed that the performance decreases, this is mainly due to the fact that LSTM can not aggregate knowledge from the repeated fixation embeddings. The LSTM module is designed to capture temporal patterns in sequential frames, which is not the case when gaze embeddings are the same. On the contrary, moving towards the lower level of discrimination abstraction (gaze-point level), the performance is improved significantly. This could be explained by the fact that the number of gaze-points is by far larger than the number of fixations (i.e. enabling LSTM to capture efficiently the gaze patterns). Therefore the rest of the experimental evaluation was focused on the gaze-point level of abstraction. Concerning the gaze-point level, crucial experimental parameters are considered to robustly validate the method. A critical parameter is to define which objects have been seen by the human subject. The latter is controlled by the use of a distance threshold, which denotes the maximum normalized distance, as defined in 1,

TABLE IV
OPTIMIZATION RESULTS: GP-DE - CRITICAL FACTORS

Object Size	Distance Threshold	Accuracy	Relevant Rate	Non-relevant Rate
>=0	None	**0.9476**	0.5577	**0.9584**
>=42	None	0.9026	0.7540	0.9095
>=0	1.0	0.9344	0.8226	0.9389
>=42	1.0	0.9200	**0.8720**	0.9221
>=0	2.0	0.9375	0.6971	0.9462
>=42	2.0	0.9158	0.7769	0.9221

that an object center must have from a gaze-point so as to be considered as seen. Threshold values equal to 1.0 and 2.0 were chosen for experimental evaluation based on visual inspection, since these values were shown to discriminate between seen and unseen objects in a satisfactory way. Applying thresholds 1.0 and 2.0 to the introduced dataset the 22.64% and 35.78% of objects are retained respectively.

Additionally, small objects (i.e. $S < 42$, based on experimentation) were neglected, since no robust gaze features could be estimated for such cases. Table IV provides quantitative object detection results for different combinations of parameters of "seen" and "object size" for the proposed approach GP-DE. The optimal performance, in terms of relevant rate, was achieved when, object is greater or equal to 42 and distance threshold equals to 1.0, parametric values were applied. This is mainly due to the fact that when the seen object parameter is reduced, the classification is restricted to the actual objects the user has seen, while when the object size parameter is increased, the least likely to be observed (i.e. smaller objects) are neglected.

As above-mentioned in Section II most of the state-of-the-art studies [10], [11], [15], [21] refer to the image level using statistical gaze features and application specific ones. Due to the limited number of available saccades in the introduced dataset, the latter consider objects as images, taking into account fixation-based features only. Although the state-of-the-art methods significantly contribute to their domain of interest (e.g. image retrieval), they fail in the relevance assessment task using the introduced dataset, due to the large number of objects, with most of them being relatively small and highly occluded. Additionally, the aforementioned methods are based on handcrafted statistical features that in combination with naive classifiers are prone to under-fitting when the classification space is complex. On the contrary, the proposed approach reveals the underlying temporal gaze patterns, in a more sophisticated manner, taking advantage of the introduced dataset, while also demonstrating outstanding generalization capabilities.

VI. CONCLUSIONS

The problem of gaze-based object-level relevance assessment prediction was addressed. In particular, a large-scale gaze annotated dataset, oriented to the visual search paradigm, was introduced for training DL architectures. Additionally, a novel gaze modeling method based on distances to objects in an image was proposed, able to handle sensor error and to capture the image context. Moreover, an LSTM-based scheme was proposed for analyzing gaze patterns, to further assess object relevancy, given a search target. The proposed method was experimentally evaluated using the introduced dataset, and it was proved to be able to predict objects' relevancy by capturing complex gaze patterns. Future work includes the investigation of including additional human responses (e.g. emotions) and their efficient incorporation in the visual analysis loop.

ACKNOWLEDGMENT

The work presented in this paper was supported by the European Commission under contract H2020-787061 ANITA.

REFERENCES

[1] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014.

[2] A. Borji and L. Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015.

[3] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178, 2015.

[4] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark. http://saliency.mit.edu/.

[5] U. Engelke and P. Le Callet. Perceived interest and overt visual attention in natural images. *Signal Processing: Image Communication*, 39:386–404, 2015.

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *CVPR*, 2015.

[8] M. Jiang, J. Xu, and Q. Zhao. Saliency in crowd. In *European Conference on Computer Vision*, pages 17–32. Springer, 2014.

[9] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009.

[10] M. Kandemir, V.-M. Saarinen, and S. Kaski. Inferring object relevance from gaze in dynamic scenes. In *ETRA*, pages 105–108, 2010.

[11] L. Kozma, A. Klami, and S. Kaski. Gazir: Gaze-based zooming interface for image retrieval. In *ICMI-MLMI*, pages 305–312, 2009.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[14] myGaze. myGaze. http://www.mygaze.com/products/mygaze-eye-tracker/, 2019.

[15] G. T. Papadopoulos, K. C. Apostolakis, and P. Daras. Gaze-based relevance feedback for realizing region-based image retrieval. *IEEE Transactions on Multimedia*, 16(2):440–454, 2014.

[16] Y. Rai and P. Le Callet. Visual attention, visual salience, and perceived interest in multimedia applications. In *Academic Press Library in Signal Processing, Volume 6*, pages 113–161. Elsevier, 2018.

[17] Y. Rai, P. Le Callet, and G. Cheung. Quantifying the relation between perceived interest and visual salience during free viewing using trellis based optimization. In *IVMSP*, pages 1–5, 2016.

[18] S. Thermos, G. T. Papadopoulos, P. Daras, and G. Potamianos. Deep affordance-grounded sensorimotor object recognition. In *CVPR*, pages 6167–6175, 2017.

[19] X. Wang, N. Thome, and M. Cord. Gaze latent support vector machine for image classification improved by weakly supervised region selection. *Pattern Recognition*, 72:59–71, 2017.

[20] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014.

[21] H. Zhang, T. Ruokolainen, J. Laaksonen, C. Hochleitner, and R. Traunmüller. Gaze- and speech-enhanced content-based image retrieval in image tagging. In T. Honkela, W. Duch, M. Girolami, and S. Kaski, editors, *ICANN*, pages 373–380, 2011.

[22] Y. Zhou, J. Wang, and Z. Chi. Content-based image retrieval based on eye-tracking. In *COGAIN*, pages 9:1–9:7, 2018.