How Old Do You Look? Inferring Your Age From Your Gaze

Tianyi Zhang, Olivier Le Meur

HOW OLD DO YOU LOOK?
INFERRING YOUR AGE FROM YOUR GAZE

A. Tianyi Zhang

Nanjing Univ. of Aeronautics and Astronautics
Department of Automation Engineering
29 Yudao Road NANJING CHINA

B. Olivier Le Meur

Univ Rennes, CNRS, IRISA
UMR 6074
F-35000 RENNES FRANCE

ABSTRACT

The visual exploration of a scene, represented by a visual scanpath, depends on a number of factors. Among them, the age of the observer plays a significant role. For instance, young kids are making shorter saccades and longer fixations than adults. In the light of these observations, we propose a new method for inferring the age of the observer from its scanpath. The proposed method is based on a 1D CNN network which is trained by real eye tracking data collected on five age groups. In order to boost the performance, the training dataset is augmented by predicting a high number of scanpaths thanks to the use of an age-dependent computational saccadic model. The proposed method brings a new momentum in this field not only by significantly outperforming existing method but also by being robust to noise and data erasure.

Index Terms— age inference, scanpath, deep network.

1. INTRODUCTION

Since we cannot process all information of our visual field, our brain uses a mechanism of selection aiming to prioritize visual chunks of the input stimulus. This mechanism, called selective visual attention, allows us to focus our visual processing resources on relevant visual information. The covert attention is a particular form of visual attention, that allows us to pay attention toward a specific location of our visual field without moving the eyes [1]. In contrast to covert attention, overt attention relies on eye movements to move our focus of attention (i.e. overt orienting). Overt attention, through eye movements, is an exogenous manifestation of the underlying cognitive processes that may or may not occur.

Eye movements are mainly composed by fixations and saccades. Fixations aim to bring objects of interest onto the fovea, where the visual acuity is maximum. Saccades are ballistic changes in eye position, allowing to jump from one position to another. Visual information extraction essentially takes place during the fixation period. The sequence of fixations and saccades is called a visual scanpath [2]. When looking at natural scenes, saccades of small amplitudes are far more numerous than long saccades [3]. In addition, horizontal saccades (leftwards or rightwards) are more frequent than vertical ones, which are much more frequent than oblique ones. These viewing biases, indicating how our gaze moves within a scene, are not systematic. They depend on a number of factors such as the scene content [4, 5], our age [6], our gender [7], and whether we suffer from a visual disease or not [8, 9]. Regarding the age influence of gaze deployment, recent studies give evidence that there exists age-related differences in viewing patterns while free-viewing visual scene on screen [10, 11]. For instance, fixation durations decrease and saccade amplitudes increase with age. These behavioral differences are due to several factors, such as the eye metamorphosis, the cognition, etc.

In this study, we designed a new method for inferring the age of an observer by only considering his visual scanpath (i.e. without considering the visual content). Five age groups are considered: 2 y.o., 4-6 y.o., 6-8 y.o., 8-10 y.o. and adults [10]. As far as we know, there exist very few methods for inferring the age of an observer from his scanpaths. Recently, Le Meur et al. [12] presented evidence that simple scanpaths-based features, such as the fixation duration and saccade length can be used to infer their age. They proposed to learn a direct non-linear mapping between simple features extracted from the ordered sequence of fixations and saccades and observer's age. They used the multi-class Gentle AdaBoost algorithm [13] to perform the classification. Good performances were reported for a binary classification involving 2 classes, i.e. 2 y.-o. versus adult groups.

The contributions of the proposed study are twofold. First, rather than extracting handcrafted gaze-based features, a 1D convolution neural network is trained to predict observer's age. Second contribution concerns the training dataset. As the amount of data for the classification is relatively small, we use a generative model for producing visual scanpaths for the different age groups. Although that these scanpaths are estimated and do not reflect exactly the actual gaze deployment of observers, it turns out that the classification performance dramatically increases, outperforming existing method in a significant manner.

This paper is organized as follows. Section 2 elaborates on the proposed scanpath-based age inference approach. Section 3 presents the performance of the proposed method. Section 4 concludes the paper.

2. DEEP INFERENCE OF OBSERVER' AGE

In this section, we present the proposed approach. Similarly to [12], the method aims to infer the age of an observer from his visual scanpath, and regardless of the visual content. Five age groups are defined, i.e. 2 y.o., 4-6 y.o., 6-8 y.o., 8-10 y.o. and adults.

2.1. Proposed model

Figure 1 presents the architecture of the proposed method, called *AgeNet*. *AgeNet* is based on a one-dimension convolutional neural network (1D CNN). It aims to learn deep features of scanpath for inferring observers' age. 1D CNN has been widely used in the field of speech recognition and sound classification [14, 15]. Like audio signals, a scanpath is an ordered time-stamped sequence. The

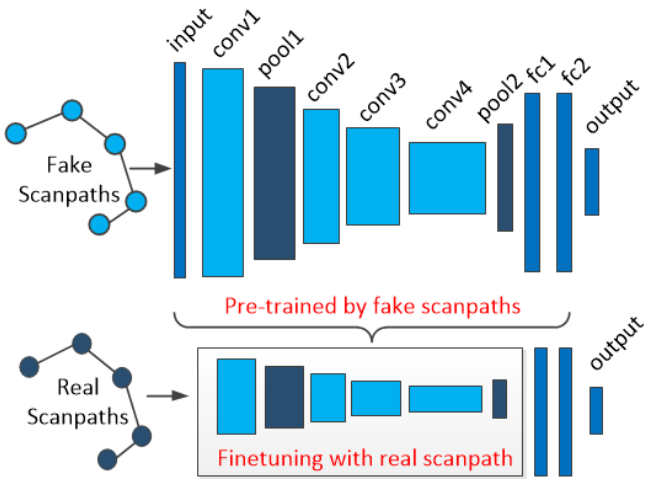

Fig. 1. Architecture of the proposed 1D CNN

Table 1. Architecture of the proposed AgeNet network

Layer	Type	#C	Kernel	Drop	Output
input	/	1	/	/	$(150, 2)$
conv1	Causal	4	$(7, 2)$	/	$(150, 4)$
pool1	Max	/	$(3, 2)$	/	$(50, 4)$
conv2	Causal	5	$(5, 4)$	/	$(50, 5)$
conv3	Causal	16	$(3, 5)$	/	$(50, 16)$
conv4	Causal	32	$(3, 16)$	/	$(50, 32)$
pool2	GlobalMax	/	/	/	32
fc1	/	/	/	0.2	128
fc2	/	/	/	/	128
output	Softmax	/	/	/	5

temporal order of scanpath is fundamental. Just after the stimulus onset, the gaze deployment is mainly stimulus-driven when top-down influences are the weakest. These influences tend to increase with the viewing time, and is likely dependent on observers' cognition [16, 17, 18]. However, unlike audio signal which often contains thousands of features in one sequence, we are considering in this study a rather short scanpath, that would be composed of less than twenty fixation points. Considering the temporal order and the length of scanpath, *AgeNet* uses 4 causal convolutional layers, as proposed in [14]. They aim to extract deep features from the input scanpath. The causal convolutional layers make sure the model cannot violate the temporal order of scanpath. That is to say, the prediction generated by the model at time t cannot be affected by any of the subsequent gaze data.

Table 1 summarizes the configuration of the proposed network. The first convolutional layer applies causal convolutions with the kernel size of 7×2 (7 represents the number of data taken into account by the convolution whereas 2 indicates that we are using the x and y coordinates of the visual fixation). Compared with Wavenet [14] and Soundnet [15], the convolutional kernel in the first layer of our network is designed to be small. Early convolutional layers aim to extract local information of scanpath while the deeper layers merge local information to get a more general description of the scanpaths. For the second, third and fourth convolutional layers, we apply causal convolutions with the kernel size of 5×4, 3×5 and 3×16, respectively. Again, the sizes of kernel are rather small for taking into account the rather small scanpath length. In addition, the use of small kernels allows us to get wider range of local clues. We did not deeper our network further since the performance did not get better by adding more layers. The max pooling layer between the first and the second convolution is set to decrease the dimension of feature maps for further convolution. After 4 layers of convolution, a GlobalMax pooling layer is added to generate a 32-dimensional feature vector. At last, the feature vector is fed to two fully-connected layers with 128 neurons each followed by an output layer composed by 5 neurons. The dropout rate for the first fully-connected layer is set to be 0.2 to avoid over-fitting.

2.2. Training

For training the proposed model, we used the eye tracking dataset of [10]. 101 subjects, including 23 adults and 78 children, participated in the experiments. These subjects were divided into 5 groups: 2 year-old group (18 participants, $M = 2.16$ ($M = Mean$), $SD = 0.22$ ($SD = StandardDeviation$)), 4-6 year-old group (22 participants, $M = 4.2$, $SD = 0.42$), 6-8 year-old (18 participants, $M = 6.6$, $SD = 0.47$), 8-10 year-old group (20 participants, $M = 8.55$, $SD = 0.67$) and adult group (22 participants, $M = 28$, $SD = 4.48$). Participants were instructed to explore 30 color pictures taken from children's books for 10 seconds. More details can be found in [10].

Unfortunately, this dataset is too small to perform a reliable training. In order to increase in a significant manner the performance of the CNN, we proceed with data augmentation. In a number of studies, the training dataset can be increased by using a combination of affine transformations. However, these transformations may not be suitable to the proposed study. Indeed, we have to take care of the ordered sequences of fixation, as well as the discrepancy between gaze deployment of the different age groups.

Our data augmentation strategy consists in artificially generating *fake* scanpaths for each age group. For this purpose, we use the age-dependent saccadic model recently proposed by [6]. Saccadic models aim to predict not only where observers look at but also how they move their eyes to explore the scene. Le Meur et al. [6] demonstrated that it is possible to tailor the saccadic model in order to capture differences in gaze behaviour between age groups. Generated scanpaths share the same statistics (e.g. distribution of saccade amplitudes, distribution of saccade orientations) as the actual ones. Thanks to this generative model, we compute, for each age group, 20.000 scanpaths, each composed of 150 fixations. Per age group, we then generate 3 million fixations.

The CNN was trained using RMSProp with a batch size of 1000 scanpaths for pre-training, and 256 for fine-tuning. During the training phase, we monitor the accuracy and the loss on the validation set to avoid over-fitting problem. The learning rate was set to $\alpha = 0.001$ over the training course. Figure 2 shows learning curve of pre-training and fine-tuning for *fake* and actual scanpaths, respectively. The training with *fake* scanpaths has a slow convergence speed. However, we do not need to get very high performance for the pre-training process: the pre-trained model gets an accuracy of 62%. Unlike *fake* scanpaths, the length of real scanpaths may vary significantly from one observer to the other. To overcome this problem, we pad the real scanpaths with zeros in order to get (150×2) values as the fake scanpaths. From Figure 2, we can see the fine-tuning process converged rapidly after only a few epochs and did not cause over-fitting problems at all. We conclude that the proposed network can effectively train the scanpaths and has a good performance of

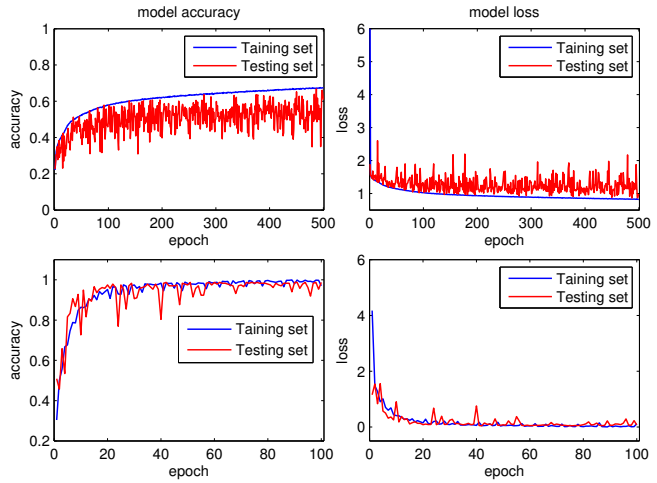

Fig. 2. Learning curves (accuracy and loss) for the proposed network trained with *fake* scanpaths (top) and by fine-tuning it with actual scanpaths (bottom).

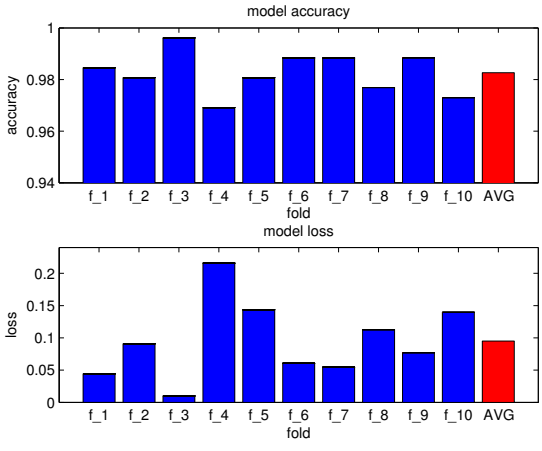

Fig. 3. Accuracy (top) and loss (bottom) of the proposed method for 10-fold cross validation.

generalization.

3. PERFORMANCE

In this section, we evaluate our network from different perspectives. Firstly, we use 10-fold cross validation approach to estimate the accuracy and loss of our model. Secondly, we compare different strategies of training using different methods of data augmentation. Then, we evaluate the robustness of the proposed method against noise and loss. At last, we compare our method with [12].

3.1. Cross validation

To evaluate the performance of our model, we cross-checked the network with 10-fold cross validation approach. The original data set was randomly partitioned into 10 equal sized subsets for training and testing. A single subset was used for testing while the remaining 9 subsets were used for training. The training and testing processes are repeated 10 times (i.e f_1,...,f_10) with each of the 10 subsets used exactly once as testing set. The purpose of this validation is to evaluate the generalization error of our model.

Figure 3 presents the result of the 10-fold cross validation. Our model (pre-trained with *fake* scanpaths and fine-tuned with actual scanpaths) achieves 98.27% and 0.095 for average accuracy and loss respectively, similar to the accuracy (98.40%) and loss (0.071) obtained when we randomly choose 90% of the data for training and 10% for testing (see the right-hand side of Figure 4). This result demonstrates that the accuracy and loss of our model do not rely on the selection of training data, which shows a very good generalization performance of the network.

3.2. Strategies of training

In this section, we compared different strategies of training using different methods of data augmentation.

Original data. We trained the network through the original data. Since the length of real scanpaths may vary significantly from one observer to the other, while the duration of them is roughly strict to 10 seconds, we sampled each scanpath every 10 milliseconds to unify their length to be (1000×2). As shown by Figure 4 (Left), the accuracy achieves 62.6%. We observed that over-fitting problem occurs after about 100 epochs. As a result of this, the accuracy could not been increased by further training.

Simple data augmentation. To overcome the aforementioned problem, we flipped and mirrored the sampled data to increase training set. Suppose a given scanpath $sp = \{(x_k, y_k)\}_{k=\{1,\ldots,1000\}}$, where (x_k, y_k) represents the spatial coordinates of the k^{th} visual fixations in the current image, we calculate the flipped scanpaths sp_f as well as the mirrored scanpaths sp_m by flipping and mirroring the spatial coordinates of visual fixations, while keeping the same temporal order. The procedure is described below:

$$\begin{aligned} sp_f &= \{(x_k, H - y_k)\}_{k=\{1,\ldots,1000\}} \\ sp_m &= \{(W - x_k, y_k)\}_{k=\{1,\ldots,1000\}} \end{aligned} \quad (1)$$

where $[W, H] = [1024, 768]$ is the width and the height of the tested images. By flipping and mirroring the scanpaths, the training set is 3-times larger than original one.

Thanks to this augmented dataset, we trained and fine-tuned our network by the augmented and original data, respectively. The accuracy was promoted to 68.6% (see Figure 4 (center)). Misclassification occurs for the groups of 6-8 y.o. and 8-10 y.o.

Pre-trained the network with fake scanpaths. As explained previously, we used an age-dependent saccadic model to generate *fake* scanpaths for each age group. The pre-training of the network with this new dataset and its fine-tuning by actual eye tracking data significantly increases the performance of the proposed model: 98.4% of the predictions are correct and only 1.6% are wrong classification. These promising results suggest that our network can extract abundant deep features from *fake* scanpaths and greatly improved the accuracy of classification by fine-tuning the network with real scanpaths.

3.3. Performance with noise and loss

We evaluated the robustness of the proposed method to corrupted data, that may arise from failure or malfunctioning of the eye tracker device. For this purpose, we add artificial noise to the spatial coordinates of the visual fixations and we remove some visual fixations

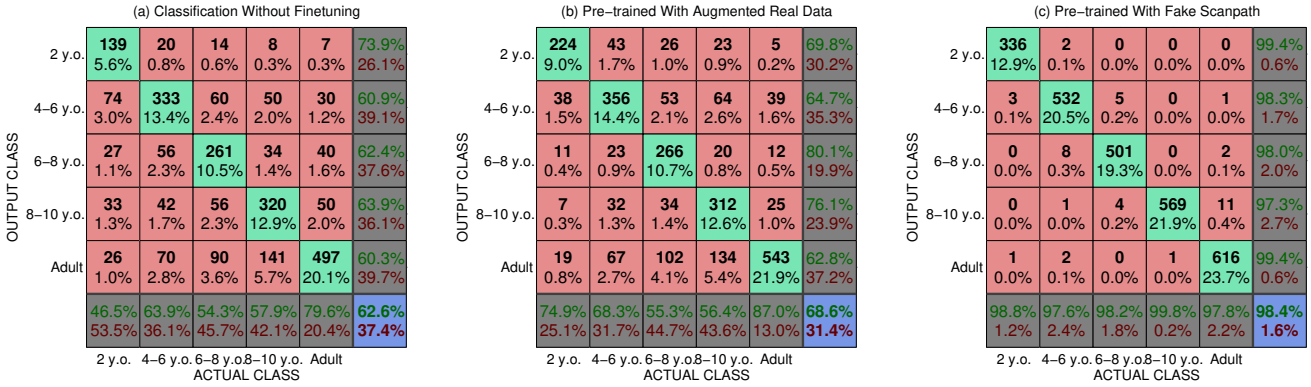

Fig. 4. Confusion matrix for the classification. (a) Trained with original data after sampling. (b) Pre-trained by data with flipping and mirroring, and fine-tuned with original data after sampling. (c) Pre-trained by *fake* scanpaths and fine-tuned by original data after zero padding.

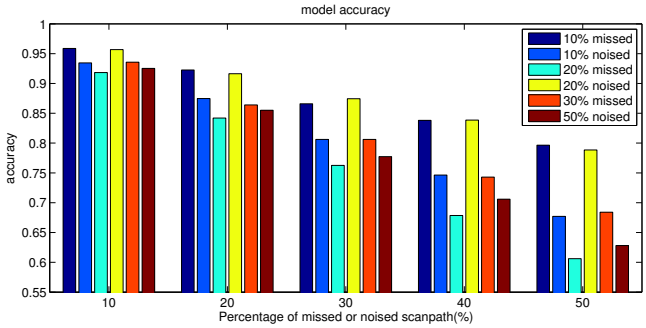

Fig. 5. Performance of the performance when input scanpaths are corrupted with noise and loss.

from the scanpaths. The question is to what extent this artificial corruption impacts the classification accuracy.

When we simulate the noisy data, $f\%$ fixation points in $p\%$ scanpaths were chosen and replaced by random values ranging from (0,0) to (1024,768). Similarly, the loss data is generated by randomly dropout $f\%$ fixation points in $p\%$ scanpaths. In our experiment, the percentage of noisy fixations and loss fixations varies between $10\%, \ldots, 30\%$, and $10\%, \ldots, 50\%$, respectively. As shown in Figure 5, when the amount of corruption is the most important, i.e. 30% of fixation points are corrupted for 50% of the scanpaths, our network still has an accuracy up to 60%. We observe that the accuracy decreases significantly when the percentage of fixation points in one scanpath are noised or missed. This is due to the corruption that breaks the temporal order of the visual fixations.

Overall, the results are promising. Indeed, in spite of noise and losses, our method performs quite well, suggesting that it could be used for predicting the observer's age when low-cost eye tracker tracks the observer's gaze.

3.4. Comparison with [12]

In this section, we compared the proposed method with [12], which predicts the age of an observer by using handcrafted features (i.e. 70 scanpaths-based features) and Random Forest. Table 2 indicates that the results of our deep learning method dramatically outperforms Le Meur et al.'s method [12], whatever the configuration. The highest improvement is observed for 4-class classification. The better performance can be explained at least by two reasons. Firstly, the deep features extracted by the network are more robust, more relevant and more numerous than the handcrafted features proposed in [12]. Secondly, the generated *fake* scanpaths provide abundant data for training, leading to an efficient training.

Table 2. Accuracy of the proposed method (%) compared to [12].

	Le Meur et al. [12]	Proposed
2-class-1[1]	91.3	99.2
2-class-2[2]	75.6	99.0
4-class[3]	52.2	98.1
5-class[4]	–	98.4

1 binary classifications between 2 y.o. and adult groups.

2 binary classifications between children ($<$ 10 y.o.) and adult groups.

3 four-class classification for 2, 4-6, 6-10 y.o. and adult groups.

4 five-class classification for 2, 4-6, 6-8, 8-10 y.o. and adult groups.

4. CONCLUSION

The proposed deep network predicts the age of an observer by scrutinizing his visual scanpath. Compared to previous method, this method increases dramatically the performance creating a new momentum in the field. Beyond its ability to predict successfully age, the method is robust to noise and losses. Although a comprehensive analysis is required to confirm these properties, they may play an important role when eye tracking data are collected in an unconstrained experimental environment and/or with low-cost eye tracker.

In future works, we want to push forward the limits of the proposed method by considering new age categories. This kind of fine-grained classification is extremely challenging because of the large intra-category variations and small inter-category variations.

5. REFERENCES

[1] Michael I Posner, "Orienting of attention," *Quarterly journal of experimental psychology*, vol. 32, no. 1, pp. 3–25, 1980.

[2] Claudio M Privitera, "The scanpath theory: its definition and later developments," 2006, vol. 6057, pp. 6057 – 6057 – 5.

[3] Olivier Le Meur and Zhi Liu, "Saccadic model of eye movements for free-viewing condition," *Vision research*, vol. 116, pp. 152–164, 2015.

[4] Olivier Le Meur and Antoine Coutrot, "Introducing context-dependent and spatially-variant viewing biases in saccadic models," *Vision Research*, vol. 121, pp. 72–84, 2016.

[5] Kévin Bannier, Eakta Jain, and Olivier Le Meur, "Deepcomics: Saliency estimation for comics," in *ACM Symposium on Eye Tracking Research and Applications*, 2017.

[6] Olivier Le Meur, Antoine Coutrot, Zhi Liu, Pia Rämä, Adrien Le Roch, and Andrea Helo, "Visual attention saccadic models learn to emulate gaze patterns from childhood to adulthood," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4777–4789, 2017.

[7] Antoine Coutrot, Nicola Binetti, Charlotte Harrison, Isabelle Mareschal, and Alan Johnston, "Face exploration dynamics differentiate men and women," *Journal of vision*, vol. 16, no. 14, pp. 16–16, 2016.

[8] Laurent Itti, "New eye-tracking techniques may revolutionize mental health screening," *Neuron*, vol. 88, no. 3, pp. 442–444, 2015.

[9] Ming Jiang and Qi Zhao, "Learning visual attention to identify people with autism spectrum disorder," October 2017.

[10] Andrea Helo, Sebastian Pannasch, Louah Sirri, and Pia Rämä, "The maturation of eye movement behavior: Scene viewing characteristics in children and adults," *Vision research*, vol. 103, pp. 83–91, 2014.

[11] Heather L Kirkorian and Daniel R Anderson, "Anticipatory eye movements while watching continuous action across shots in video sequences: A developmental study," *Child development*, vol. 88, no. 4, pp. 1284–1301, 2017.

[12] Olivier Le Meur, Antoine Coutrot, Zhi Liu, Pia Rämä, Adrien Le Roch, and Andrea Helo, "Your gaze betrays your age," in *Signal Processing Conference (EUSIPCO), 2017 25th European*. IEEE, 2017, pp. 1892–1896.

[13] Yoav Freund and Robert E Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*. Springer, 1995, pp. 23–37.

[14] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[15] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016, pp. 892–900.

[16] Derrick Parkhurst, Klinton Law, and Ernst Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision research*, vol. 42, no. 1, pp. 107–123, 2002.

[17] B.W. Tatler, R. J. Baddeley, and I.D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision Research*, vol. 45, pp. 643–659, 2005.

[18] Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau, "A coherent computational approach to model the bottom-up visual attention," *IEEE Trans. On PAMI*, vol. 28, no. 5, pp. 802–817, May 2006.