NIPS2017
Neural Information Processing Systems 2017
Dec 4 , 2017 - Dec 9 , 2017, Long Beach, United States

Reviews For Paper

Track	Blue
Paper ID	263
Title	Predict Where Human to Look in Panoramic Video: A Reinforcement Learning Approach

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Overall rating:	Marginally below acceptance threshold
Confidence score:	The reviewer is confident but not absolutely certain that the evaluation is correct
Detailed Comments:	The paper presents a reinforcement learning approach to estimate saliency maps for panoramic videos. Head movement scan-paths are modeled as actions of an agent in the RL framework. The goal is to optimize the reward in predicting directions and magnitude of the scan-paths. A new dataset is introduced for the task. Strengths: + Along with the recent papers [20,21], this work is setting a new interesting computer vision direction aboutpanoramic videos + The analysis of the dataset and findings presented in section 2.2 is valuable. It is interesting to see that some intuitive findings are also confirmed by the data. One comment on that is: Table 1 shows the CC between two groups, that what would be the CC of a group compared to random Gaussian maps or uniform maps? This would set the baseline for Table 1. Weaknesses: - Although working on panoramic videos is novel, it is not clear how estimating saliency maps might benefit any application. Many works in past showed that saliency for images and videos can be combined or even learned jointly with other tasks (e.g., object recognition, object detection, activity localization, ...). However, this work does not evaluate that the predicted saliency is valuable for improving other tasks. - One of the main issues of this paper is the assumption that the head position can approximate the eye fixation of the subject. Although, this might be true in some cases, but it is also true that subjects can move the eyes and therefore head and eye directions are not aligned. Therefore the head position might be not the best information to estimate. It is not clear how this point would influence the algorithm, since it is not clear how the saliency maps can be used in some applications. It can be that for some applications the head position is enough, but in many others it is not. I would have expected a discussion about this point and also experiments that uses the predicted saliency map on other application(s). - The propositions in Sec. 3.1 are interesting, but not rigorous. A formal math definition of them would have been appreciated. Also, the proof of proposition 2 in the supplementary material is not so straightforward. In particular, what is the distribution "subjects" in Eq. 5 and how Eq. 6 is derived. In general, the formalism of the proofs needs to be improved. - Some decision about the method are not justified. -- During training the reward is estimated using the ground-truth scan-paths. However, how this is done during inference is not explained. -- Line 197: how the LSTM is connected to \nu_t and \pi_t? What is the layer after the LSTM in Fig. 6? -- It is not clear how the state value is defined (V line 222). Since this is probably inherited by the Q-learning framework, V should have been defined in

	the context of this paper. -- It is not clear why the magnitude and the direction are decoupled: the magnitude is estimated directly by the network, while the direction is estimated by the policy and RL. It could be possible to have a policy for both direction and magnitude and then estimate directly the reward in Eq. 3 without considering the one in Eq. 2 (since it is already part of Eq. 3). This choice would have required a better justification and discussion. -- The set of experiences in line 208 contains o_t and f_{t-1}. It is not clear why f_{t} and/or o_{t-1} are not used. - The experiments are interesting however limited. -- As mentioned above, saliency prediction might be valuable as long as it is attached to an application. The authors should have carried out experiments on an application that might benefit from using the saliency maps, e.g., object detection in panoramic videos, or compression of videos. -- The parameters are optimized in the test set (line 239). This is not a good practice for ML applications. The authors did the same with BMS and OBDL? If not, the comparison is unfair. -- It is not clear how BMS and OBDL were used. Do they use the entire video and frames to estimate the saliency map? -- The information about the length of the scan-paths is missing (estimated one fixation per frame or more?) Neutral: ~ The authors could have used the dataset presented already in [21] for the experiments. However, since the paper is very recent, that dataset might have not been available at the moment of the submission.

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
Overall rating:	Marginally above acceptance threshold
Confidence score:	The reviewer is confident but not absolutely certain that the evaluation is correct
Detailed Comments:	The interesting part of this work is the head movement database in panoramic video, but the description of the findings and the description of the method to make the predictions are both quite fuzzy (i.e. not the expected quality for NIPS). For instance, the architecture of the predictor is not justified: why a cascade of convolutional layers plus a long-short-term memory layer? More examples: * Concepts involved in findings 2-4 are confuse. What is magnitude and direction of the scan-paths? The explanation in "Finding 2" only says that scan-paths can be "decomposed" in magnitude and direction. How?. In finding 4, direction means the way the observers goes from one point to the next?. I only figured out the magnitude-direction decomposition after reading the supplementary material and looking at fig5. * What is the role of different workflows in Fig5? * Proof of proposition 1 (in the supplementary material) is not a proof, but the description of a method. Methodological questions about the database: (1) the initial location of the observers in each sequence is the same? (2) Does inter-subject correlation depends on time?, i.e. observers converge to look at certain location *after* some exploration time? (3) Do the scenes have a markedly anisotropic 3D nature (e.g. corridors, action in a single direction)... Minor comments: Fig 1 is not informative: heat maps are simply a sum of Gaussians? How do you choose the width of the Gaussians. Symbols and text in figures is too small.

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
Overall rating:	Ok but not good enough - rejection
Confidence score:	The reviewer is absolutely certain that the evaluation is correct and very familiar with the relevant literature
Detailed Comments:	The paper presents a new dataset for head movements (HM) in panoramic video, and proposes a deep reinforcement learning approach to predict the head movement maps in such videos. The dataset itself could be valuable to the community, being the first publicly available dataset for panoramic video with head movements. However, the part of the paper using deep reinforcement learning to predict HM is lacking details and is not well investigated. The main issue is that the model is never introduced clearly nor explained well. The authors just refer to Fig 6 to introduce the model. Fig 6 does not contain any annotations to mark the variables used in Sec 3.2. What does HM speed mean? Speed is never talked about in the text. It is also not clear why the DRL network is predicting the value function V, when the policy is also directly being predicted. How exactly is the value function defined in terms of the reward function defined in Eq.3? Further, no details of the conv and LSTM layers are provided. Was the entire network trained from scratch? was it initialized using pretraining on images? Other points: * In table 1, the CC results should also be shown for a control such as FCB. * In Finding 2, how does the magnitude of HM scan paths compare with average head movement? Is it likely that the average HM scan path is independent of the content? * In the DRL framework, it looks like the number of time steps tmax should be known ahead of time. Is this fixed for all videos? How is this determined for a new video? * have the authors considered using a soft attention model such as DRAW: A Recurrent Neural Network For Image Generation, Gregor et al., ICML 2015 (https://arxiv.org/pdf/1502.04623.pdf). This might allow the prediction of locations without recourse to reinforcement learning. Minor comments: Line 29 should be "there exists few works" The FOV size should be mentioned in Line 41-55