Study of Static and Dynamic Human Visual Saliency

April 29, 2022

Abstract

The study of human visual saliency in images and videos along with the corresponding gaze information has been a great research interest for the past few years. Particularly, egocentric videos and gaze data contain a treasure of information about human understanding of the stimuli, story about wearer's perception, etc. However, there are a very few gaze datasets collected in a free viewing style with end-to-end processing pipeline. In this study, a new egocentric vision dataset is collected from human participants (with normal vision) in a free-viewing style at a technological museum in IIT Kharagpur. The gaze and video streams are post-processed for handling incoherent information and head distortions, respectively. Considering the viewing behavior of all the participants towards an exhibit, a representative scanpath has been built by generalizing multiple viewers' gazing styles. The processed egocentric gaze dataset has been used for evaluating a newly proposed visual feature and gaze driven egocentric video retargeting approach. The approach involves grouping video frames depending on the visual content, then generating a panning window followed by zooming in and out during the scene and transition frames, respectively. Further, a statistical learning model based on the state space analysis has been proposed for extracting the inherent gaze data from gaze dynamics containing information about the kind of visual stimuli. The model is optimized for their parameters and is used for classifying unknown gaze information and retrieving the corresponding video. STEAS model is found giving promising performance on both the datasets when compared with other state-of-the-art (SOTA) models. For predicting visual saliency in images, a cross-concatenated multi-scale residual (CMR) block based network with efficient feature encoding has been proposed. A series of CMR blocks, a DIM block, and a newly proposed decoder, are combined to propose a network named CMRNet. A framework has also been proposed for deriving an explainable model from its corresponding deep image saliency model using human perception theories. Three SOTA deep saliency models namely UNISAL, MSI-Net, and CMRNet are used for this purpose. The explainability framework addresses human perception aspects and

efficiently interprets a deep saliency model. A great similarity between components of HVS and the base operations of the explainability framework has been established by discussing its possible anatomical substratum. As an extension to image saliency, a two-pathway deep model has been proposed that efficiently integrates spatial and temporal domain features with CMR blocks using dense residual cross connections, named TP-CMRNet. Frame and optical flow pathways extract tightly integrated features followed by a BD-LSTM module, a simple auto-encoder and a decoder. Experiments show the effectiveness of image and video saliency models and their interpretations done using explainability framework. TP-CMRNet excels in terms of parameters and inference time by outperforming all the relevant deep dynamic models.
Keywords: *Egocentric gaze dataset, egocentric video retargeting, statistical learning model, deep image and video saliency networks, explainability framework.*