Learning Gaze Transitions from Depth to Improve Video Saliency Estimation

George Leifman
Amazon
gleifman@amazon.com

Dmitry Rudoy
Intel Corporation
dmitry.rudoy@gmail.com

Tristan Swedish
MIT Media Lab
tswedish@mit.edu

Eduardo Bayro-Corrochano
CINVESTAV
edb@gdl.cinvestav.mx

Ramesh Raskar
MIT Media Lab
raskar@mit.edu

Abstract

In this paper we introduce a novel Depth-Aware Video Saliency approach to predict human focus of attention when viewing videos that contain a depth map (RGBD) on a 2D screen. Saliency estimation in this scenario is highly important since in the near future 3D video content will be easily acquired yet hard to display. Despite considerable progress in 3D display technologies, most are still expensive and require special glasses for viewing, so RGBD content is primarily viewed on 2D screens, removing the depth channel from the final viewing experience. We train a generative convolutional neural network that predicts the 2D viewing saliency map for a given frame using the RGBD pixel values and previous fixation estimates in the video. To evaluate the performance of our approach, we present a new comprehensive database of 2D viewing eye-fixation ground-truth for RGBD videos. Our experiments indicate that it is beneficial to integrate depth into video saliency estimates for content that is viewed on a 2D display. We demonstrate that our approach outperforms state-of-the-art methods for video saliency, achieving 15% relative improvement.

1. Introduction

In recent years we have witnessed a dramatic improvement of 3D-capable acquisition equipment; 3D cameras, *e.g.* Kinect and RealSense, have become highly popular and affordable. Moreover, in the near future many laptops and tablets are expected to be shipped with integrated 3D cameras. We also see a considerable progress in 3D display technologies. However, high-quality 3D displays are still expensive and not easily accessible to the average consumer. Combination of the factors above leads to a world where the 3D content is easy to acquire but hard to display. For these reasons, we explore the problem of predicting the human foci of attention when viewing content that contains
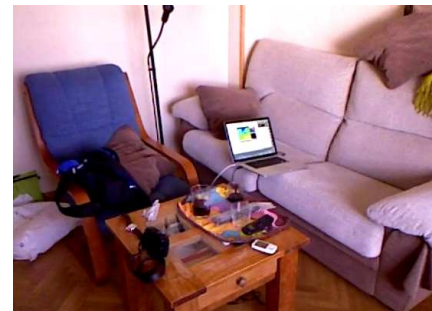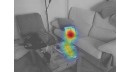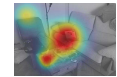

Figure 1. Our depth-aware video saliency is more similar to the ground-truth than the state-of-the-art method [34].

a depth map on regular 2D screens.

Saliency detection in video sequences has attracted a lot of attention in recent years because of its contribution for various computer vision applications, which include segmentation, classification, key-frame selection, retargeting and compression. 3D visual information supplies a powerful cue for saliency analysis. This has been shown by numerous studies that investigate the effect of depth information for image and video saliency [3, 17, 24, 30, 31]. The eye movement patterns in 3D stereoscopic movies have been investigated as well [13] and were proven to differ from the eye movement when viewing the same content on a 2D screen. This difference is beyond the scope of this paper. We focus on the scenarios where depth information exists but is not displayed to the viewer.

We propose a novel Depth-Aware Video Saliency approach that exploits depth information to establish saliency in video sequences (Figure 1). Since depth influence on the saliency is not clear, integrating such information into video saliency is not as simple as adding a prior. Figure 2 demonstrates that sometimes the closest object attracts the most attention, while sometimes distant objects are the salient ones.

To determine the correct impact of depth on saliency, we

(a) close = salient   (b) distant = salient

Figure 2. **An ambiguous impact of depth on saliency.** In some cases, the closest object is the salient one (a). In other cases, the fact that the object is distant increases its saliency (b).

train a generative convolutional neural network. The network predicts a saliency map for a frame, given the estimated map of the previous frame. This prediction resolves the ambiguity of depth impact by learning its influence on the saliency.

To the best of our knowledge, a comprehensive eye tracking database for video sequences containing depth information is yet to be developed. To evaluate the performance of our approach we introduce the Depth-Aware Video Saliency dataset. This dataset is focused on unedited videos, where the viewers' attention is not altered by a human editor. Ground-truth was established by recording eye-fixations while viewing the video on regular screens, ignoring the depth channel. To establish an objective baseline for the comparison we incorporate depth into the video saliency approach recently proposed by [34].

In this paper we show how to get solid performance improvement for video saliency estimation when depth data is available along with the RGB frames. We claim that in many situations depth information already exists or can be easily obtained. For example, many robots or autonomous vehicles already have depth sensor; recent motion pictures are shot in 3D; many video conference settings have multiple cameras and depth can be easily estimated. In these cases the information might be consumed on an ordinary 2D screen and our approach significantly improves saliency estimation results with very minimal investment.

Our contribution is threefold.

- First, we introduce a novel depth-aware video saliency approach and implement it using a generative convolutional neural network. We show that our approach outperforms state-of-the-art methods for video saliency.

- Second, we present a new comprehensive dataset of RGBD videos with eye-fixation ground-truth.

- Third, we experimentally demonstrate that learning-based integration of depth information into a saliency estimation framework improves its accuracy.

The rest of the paper is organized as follows. Section 2 reviews the previous work. Section 3 describes our database and the baseline algorithm for depth-aware saliency. Section 4 introduces our depth-aware video saliency approach. Section 5 presents our experimental results. Section 6 concludes the paper.

2. Related Work

Researchers have studied human visual attention for decades. This section discusses two saliency aspects closely related to our research: video saliency and depth-aware saliency.

Video Saliency: Most existing motion saliency methods improve image saliency models by taking into account simple motion cues. For instance, Guo *et al.* [8] adopt an efficient method based on spectral analysis of the frequencies in the video. Similarly, Cui *et al.* [4] concentrate on motion saliency only by analyzing the Fourier spectrum of the video along X-T and Y-T planes. Mahadevan and Vasconcelos [27] model video patches as dynamic textures, to handle complicated backgrounds and a moving camera. Seo and Milanfar [36] propose using self-resemblance in static and space-time saliency detection. Hou and Zhang [12] propose using incremental coding length to measure the rarity of features. Zhong *et al.* [47] use optical flows based on the dynamic consistency of motion. Rudoy *et al.* [34] narrow their focus to a sparse set of candidate gaze locations and then use learning to predict conditional gaze transitions over time. Zhou *et al.* [48] introduce motion saliency method that combines various low-level features with region-based contrast analysis to generate low-frame-rate videos. Zhang *et al.* [46] detects spatiotemporal visual saliency based on the phase spectrum of the videos. Recently, the deep learning approach was also utilized for saliency detection [22, 42]; others tried random walks [21] and super-pixels [26].

Depth-Aware Saliency: Compared to the number of saliency papers on 2D images and 2D videos, only a small amount of work on 3D content visual attention can be found. For example, Jansen *et al.* [16] investigate the influence of disparity on viewing behavior in the observation of 2D and 3D still images. Liu *et al.* [25] examine visual features at fixated positions for stereo images with a natural content. Wang *et al.* [41] examine "depth-bias" in the task-free viewing of still stereoscopic synthetic stimuli. A review of 3D visual attention papers is presented in [40].

In our research we assume that depth information exists but is not displayed to the viewer. Thus, we are less concerned about the impact of 3D viewing experience on the human visual perception. We are interested in exploiting depth for saliency estimation, when the stimuli are two-dimensional. To the best of our knowledge there is no such

previous work for video saliency. For still images, integrating depth information into the saliency model was first proposed more than a decade ago by Ouerhani *et al.* [31]. They extend the approach of [15] and treat depth as just another channel, along with color and other cues.

The recent dramatic improvement of 3D-capable acquisition devices has prompted many researchers to find more effective ways to exploit depth for image saliency calculation. Ciptadi *et al.* [3] explicitly construct 3D layout and shape features from the depth measurements. Lang *et al.* [24] present a depth prior for saliency learned from human gaze information. This saliency prior produces a saliency map that is then either directly added or multiplied by the saliency results of other methods. A novel saliency method, which is based on an anisotropic center-surround difference, is proposed in [17]. Desingh *et al.* [5] verify that depth really matters on a small dataset and propose to fuse saliency maps, produced by appearance and depth cues independently, through non-linear support vector regression. Finally, Peng *et al.* [32] propose a saliency model, where depth and appearance information from multiple layers is taken into account simultaneously, rather than simply fusing depth-induced saliency with color-produced saliency.

3. Baseline Dataset and Algorithm

Before presenting our novel depth-aware video saliency approach we discuss a baseline which is required for any evaluation. Providing a fair performance evaluation of our approach requires the following two components, which are described in the rest of this section:

1. dataset of RGBD videos containing ground-truth of human attention

2. state-of-the-art video saliency estimation algorithm, extended to take into account depth information

3.1. Depth-Aware Video Saliency Dataset

An overview of eye-tracking datasets is found in [43]. To evaluate the performance of our approach a comprehensive database containing a ground-truth of human attention on RGBD video sequences is needed. We are not aware of such a dataset. Thus, we built a new dataset of RGBD videos and captured human attention when displaying the RGB information only. This dataset will be publicly available upon publication.

Collecting the videos: The videos in our dataset should represent the scenarios where depth-aware saliency is beneficial. Thus, we focus on RGBD videos acquired by built-in phone/tablet/laptop depth/stereo cameras or 3D sensors, such as Kinect or LiDAR. We consider acquisition devices that can be either static or installed on moving vehicles or robots. Thus, we include video sequences of static and dynamic scenes, acquired by static and dynamic sensors, in-

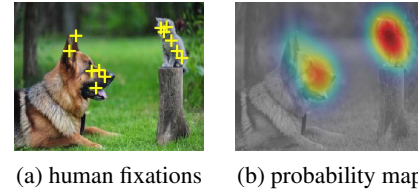(a) human fixations (b) probability map

Figure 3. **Gaze probability map.** Given a sparse ground-truth set of human fixations, marked with yellow '+' per each viewer, we convert it into a dense probability map by convolving with a constant-size Gaussian kernel (σ is 5% of the frame diagonal).

doors and outdoors. We cover scenarios such as video conference, surveillance, tracking and obstacle avoidance.

To achieve diversity, we included in our dataset RGBD videos from seven publicly available databases [20, 23, 33, 37, 38, 44, 49]. These datasets were not designed for saliency detection, but rather for other tasks, such as reconstruction, tracking or action recognition. Thus, they lack the ground-truth of human attention. We have included only videos where the color and depth frames are fully synchronized. After ignoring videos with missing regions of the depth map, we included in our dataset 54 videos with varying durations ranging from 25 to 200 seconds. The videos were converted to a 30 fps frame-rate, resulting in approximately 100K frames across all videos.

Building the ground-truth: To build a ground-truth for our dataset we conducted a comprehensive user study. To identify where participants were looking while watching the films, we monitored their eye movements using a *Gazepoint GP3* Eye Tracker[1]. Video presentation was controlled using the Gazepoint Analysis Standard software.

For the study we recruited 91 participants (52 males, 39 females). Ages ranged from 20 to 67 with the mean age of 26. All the participants had normal or corrected-to-normal vision and were naïve to the underlying purposes of the experiment.

First, we performed a calibration procedure by asking the participants to look at five red dots appearing on the screen. Then, we informed the participants that they would watch a series of short videos. We displayed the videos in random order at a viewing distance varying between 70 and 110 cm. The videos were scaled to the same resolution and displayed in full-screen. We do not use fixation duration, and the tracker uses 60Hz.

Finally, to get a dense probability map, we convolved the sparse set of fixations from all the participants with a constant-size Gaussian kernel. Figure 3 demonstrates an example of the fixation set and its resulting probability map.

Quality of the ground-truth: To assess the quality of the collected ground-truth we quantify the homogeneity of the human fixations. In other words, we measure how much the

---

[1] http://www.gazept.com/product/gazepoint-gp3-eye-tracker

fixation map "explains itself." This quality measure also serves as an upper bound for saliency prediction.

To calculate the quality, we randomly divide the set of individual fixation maps, $\mathbf{F}$, into two subsets and the probability maps of each subset are compared using χ^2 metric. We repeat this random process N times and average the results to obtain the homogeneity score for each frame:

$$Q = 1 - \frac{1}{N} \sum_{i=1}^{N} \chi^2(M(\mathbf{F}_i), M(\mathbf{F} \setminus \mathbf{F}_i)), \quad (1)$$

where $\mathbf{F}_i \subseteq \mathbf{F}$ is a random subset of the fixation set $\mathbf{F}$ in the $i - th$ iteration and $M(\mathbf{F})$ is the dense probability map of $\mathbf{F}$. The final quality score for each video is calculated by averaging the scores over all the frames.

We compare the quality of our ground-truth to quality of the DIEM (Dynamic Images and Eye Movements) dataset [29]. DIEM is a well-known dataset, which has been widely used for evaluation of video saliency techniques. It includes 84 videos of varying styles. The dataset is provided together with gaze tracks of about 50 participants per video. The video clips included in the DIEM dataset lack any depth information.

Figure 4 compares the quality of our gaze tracking ground-truth to the quality of the fixation data in the DIEM dataset. Perfect correlation, *i.e.* all the participants followed the exact same focus point on the screen, corresponds to a score of 1.

The DIEM dataset contains movies that have been professionally filmed and usually edited with a goal to attract human attention to specific objects on the screen. This is especially noticed in commercials and movie trailers. Thus, we expect high homogeneity of the human fixations; Figure 4(a) demonstrates the average score of 0.87 varying from 0.78 to 0.93 between the different movies.

Our dataset includes mostly unedited clips, filmed either by amateurs or automatically. We think that these clips better represent typical saliency use cases, since is the edited videos the viewers' attention is directed by the editor. Thus, our dataset is more "challenging" in this regard, and we cannot always expect people to agree on one specific focus of attention. Still, as shown in Figure 4(b), the average score of our dataset is 0.84 varying from 0.74 to 0.91. These comparable results indicate that most people agree on the same limited number of attention foci, even when the videos were filmed without trying to draw human attention to specific objects. We also verify visually that the viewers are not looking at a single point most of the time.

We believe that our dataset represents the wide range of common scenarios where depth-aware saliency is beneficial. The size of our dataset (54 videos) was chosen to be similar to the other two most popular datasets for video saliency: DIEM [29] and CRCNS [14], which include 85 and 50 videos, respectively.
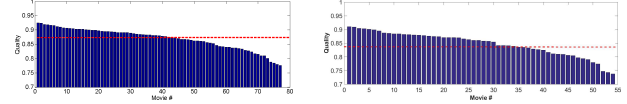
(a) DIEM dataset [29] (b) Our dataset

Figure 4. **Quality of the gaze ground-truth.** To assess the quality of the collected ground-truth we measure how much the fixation map "explains itself". (Each bar corresponds to one movie; the red line indicates the average). The quality of the fixation maps in our dataset is comparable to the one of the DIEM dataset [29].
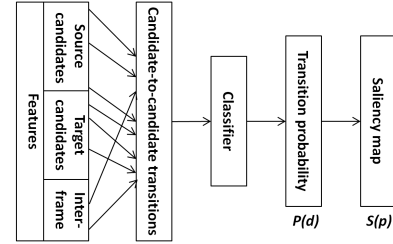

Figure 5. **Saliency estimation using explicit transition prediction.** Initially, the features are calculated on the source candidates (previous frame saliency) and the target ones (detected). The aggregated features that represent gaze transitions are fed to a trained classifier that outputs a probability of transition to target candidates. Finally, the probabilities are integrated into a saliency map.

3.2. Baseline Depth-Aware Algorithm

To establish a fair and objective baseline for the comparison we extend the algorithm recently proposed by [34] with depth information in its key stages. Let us first summarize the original scheme and then explain our extensions.

Original scheme: As demonstrated in Figure 5, first, a sparse set of candidates is generated for each frame. Then, a classifier that predicts gaze transitions between various candidates of different frames is trained. The feature space of the classifier accounts for the candidates' properties (*e.g.* saliency magnitude, motion magnitude) and also captures the relation between the candidates (*e.g.* the distance between their locations). Next, applying the trained classifier, the gaze transition probability from each candidate of a source frame to each candidate of a target frame is calculated. Finally, a saliency map is generated for each frame based on transition probabilities.

The candidate locations are generated for all video frames based on three cues. First, Graph-Based Visual Saliency (GBVS) [9] is calculated for each frame. Second, some high-level cues (*e.g.* human figures and faces) are added.Third, to account for motion the optical flow is calculated between the consecutive frames. Finally, each candidate location is represented by a Gaussian blob, calculated by applying mean-shift clustering and Gaussian fitting on the normalized saliency maps and on the differences in the optical flow magnitude.

After generating a set of candidates in each frame, the

gaze transition probability from the candidates of two consecutive frames is calculated. All the possible pairs of candidates are considered and each pair is associated with a feature vector. The feature vector consists of (1) the mean saliency of the candidate neighborhood, (2) Difference-of-Gaussians of the optical flow vectors and of their magnitude, (3) discrete candidate labels: face, body and center and (4) geometric features: the distance between the candidates and the distance from the candidate location to the center of the frame. A classifier is trained on a subset of videos based on the eye-tracking ground-truth. Finally, the transition probabilities are calculated by applying the classifier on the entire dataset.

Depth-aware extension: We incorporate depth information in three key stages: static saliency estimation, optical flow calculation and gaze transition modeling. Our experiments show that all three improvements are vital.

First, depth-aware image saliency is used for generating candidate locations. We calculate depth-aware saliency based on a multi-stage RGBD model recently proposed in [32]. This technique accounts for both depth and appearance cues derived from low-level feature contrast, mid-level region grouping, and high-level prior enhancement. **Second**, depth is used for the optical flow calculation between consecutive frames. Instead of calculating optical flow on three color channels, we use an additional channel — the depth. This way our motion estimation is more accurate than in the previous methods, especially for objects moving on a similarly colored background. We have considered implementing more sophisticated methods for dense motion estimation using color and depth (*e.g.* [10]). However, the complexity of such techniques is high, making them impractical to apply to videos. **Third**, when calculating the feature vectors associated with each candidate pair we exploit depth information, by adding a signed difference in candidates' depths to the set of the geometric features.

All the candidates in the source and destination frames are examined and labeled as positive or negative. The transitions are positive when they connect between the candidates that are aligned with the human fixations. Other transitions are marked as negative.

An SVM classifier is trained on the feature vectors and their corresponding labels. The output of the classifier is the signed distance from the separating hyper-plane. This distance is proportional to the confidence $C(s, d)$ of transition from the candidate s of the source frame to the candidate d of the destination frame. The overall probability $P(d)$ of gaze to reach the destination candidate d is calculated by combining all positively classified transitions to candidate d. Thus, ignoring transitions with negative confidence, we

calculate $P(d)$ as follows:

$$P(d) = \frac{1}{|\mathbf{N}_S|} \sum_{s \in \mathbf{N}_S} S(s) \cdot \max\left(C(s, d), 0\right), \quad (2)$$

where $\mathbf{N}_S$ is the set of all the sources and $S(s)$ is the saliency of the source candidate.

Finally, the saliency of pixel p in the destination frame is given by a sum of constant-size Gaussians around each destination candidate d, scaled up by the probability $P(d)$:

$$S(p) = \frac{1}{|\mathbf{N}_D|} \sum_{d \in \mathbf{N}_D} P(d) \cdot \exp\left(-\frac{||p - d||^2}{2\sigma^2}\right), \quad (3)$$

where $\mathbf{N}_D$ is the set of all the destination candidates in a given frame and σ equals 5% of the frame diagonal.

4. Our Approach

This section presents our approach for depth-aware video saliency estimation, which is based on the following three principles. **First**, in video the gaze usually slightly varies between frames, and when it does change significantly, it is constrained to a limited number of foci of attention. **Second**, people usually follow the action by shifting their gaze to a new interesting location. Thus, we consider a sparse candidate set of salient locations and use learning to predict transitions between them over time. **Third**, in addition to the above two principles, which are common to many previous video saliency approaches, we claim that depth perception has an impact on human attention. This claim is supported by our experimental results as shown in Section 5. Note that in some cases, the closest object attracts the most attention (Figure 2(a)) and in other cases, a distant object causes humans to concentrate their attention on it (Figure 2(b)). To resolve this ambiguity, we propose incorporating depth into the learning process.

To realize the above three principles, we propose to train a generative convolutional neural network to predict the saliency for each frame. According to the first and the second principles, the gaze transition between frames is limited to a small number of locations. Therefore, it is safe to assume that it is feasible to learn a compact representation for gaze transition between frames. As shown in Figure 6, our network's input is the saliency calculated for the previous frame and additional information from the current frame. Then the data is encoded in a compact way, which represents the gaze transition between frames and only the saliency of the next frame is reconstructed.

Work on generative models typically addresses the problem of unsupervised learning of a compressed, distributed representation (encoding) for a set of data. Such networks are usually used to generate samples from a hidden representation. The most known examples are auto-encoders
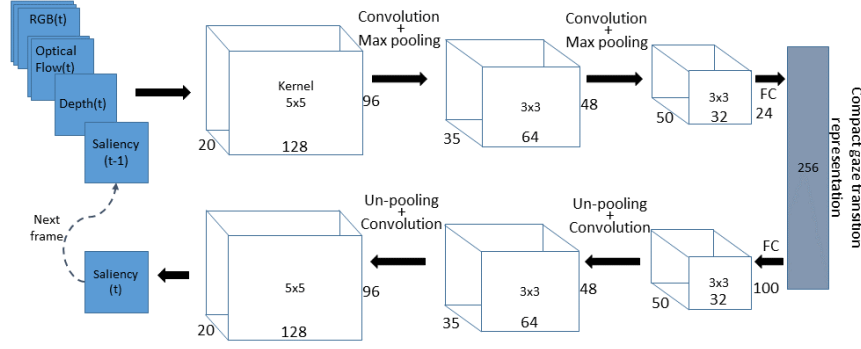
Figure 6. **Saliency reconstruction using a generative convolutional neural network.** The input is the saliency calculated for the previous frame and additional information from the current frame. Then the data is encoded and only the saliency of the current frame is reconstructed.

based on Restricted Boltzmann Machines (RBMs) [11] and Deep Boltzmann Machines (DBMs) [35]. A basic auto-encoder is an artificial neural network used for learning data coding. Following the notation from [28], first, the input x is mapped to the latent representation h using a function $h = \sigma(Wx + b)$. This compressed representation is then used to reconstruct the input by a reverse mapping of $\hat{x} = \sigma(W'h + b')$. The weights of W are optimized, minimizing an appropriate cost function over a given training set. Usually the same weights for encoding the input and the decoding are used, *i.e.* $W' = W^T$.

Conventional auto-encoders are fully connected and consequently ignore the spatial image structure. This introduces redundancy in the parameters, forcing each feature to be global. We base our architecture on convolutional auto-encoder structure [28, 1], whose weights are shared among all locations in the input, preserving spatial locality. In the aforementioned auto-encoders, an input image x is passed through the hidden layers, computing activations at all the layers to obtain the output image $\hat{x}$. Then, the deviation error from the input $e = x - \hat{x}$ is calculated and back-propagated through the network.

As shown in Figure 6, our generative convolutional neural network gets as an input a set of seven images $X = \{x_i\}_1^7$ and reconstructs only one image. For each frame, the set X consists of the following seven channels: RGB (3 images), optical flow (2 images), depth map and a saliency map $S(t-1)$ calculated for the previous frame. The output of the network is an estimation of a single saliency map $\hat{S}(t)$ for the current frame. Therefore, the deviation error is calculated as $e = S(t) - \hat{S}(t)$, introducing an asymmetry between the input and the output, *i.e.* $W' \neq W^T$. Note, $S(t)$ refers to the ground-truth map. We use χ^2 distance between two distributions of the saliency maps. Then, the error is back-propagated through the network, updating the weights using stochastic gradient descent. The whole process is recursive, where we start with a saliency map $S(0)$ which consists of a single Gaussian located in the center of

the first frame. Then the estimated saliency map $\hat{S}(1)$ is fed as an input $S(1)$ to the network for the next frame.

Finally, following our first principle, we strive to estimate a sparse set of attention foci. However, the nature of our generative network is to reconstruct relatively smooth output images. Thus, we add an output post-processing stage to sharpen the peaks of a limited number of attention foci. This is done by applying mean-shift clustering and Gaussian fitting.

Architecture details: We experimented with different network configurations and the best results are achieved by the network shown in Figure 6. First, the input 7-channel image is passed through an encoder. The encoder consists of three layers of convolutions followed by max-pooling whose sizes are 128x96, 64x48 and 32x24 with kernel sizes of 5x5, 3x3 and 3x3, respectively. Then the data is encoded in 256 latent variables fully connected to the encoder and the decoder. The decoder consists of three layers of un-pooling followed by convolution of the same sizes in reverse order. The un-pooling is performed according to the scheme proposed by [6].

We used stochastic gradient descent with a fixed momentum of 0.9. For 200 epochs the learning rate was 10^{-4} and then for an additional 200 training epochs we divided the rate by two after every 50 epochs. The network is trained on a subset of two-thirds of the videos and the training error is estimated based on the eye-tracking ground-truth.

Since our saliency learning is recursive, only frames from different videos are used simultaneously, limiting the batch size to the number of videos in the training set. In other words, the first batch consists of all the first frames, the second batch consists of all the second frames, when the input to the second batch is the saliency maps estimated in the first batch. For the simplicity of the exposition we used the term "previous frame"; however, because humans usually fixate in about 300ms, or 10 frames under common frame-rate of 30 fps, in the implementation the $S(t-1)$ is

changed to include a single saliency map of 10 frames back.

5. Results

This section presents quantitative and qualitative evaluation of our technique.

Quantitative evaluation: For quantitative evaluation we use two common metrics: area-under-curve (AUC) and χ^2 distance between distributions. AUC is the area under the Receiver Operating Characteristics (ROC) curve [2]. Human fixations are considered as the positive set, while the negative set is formed from randomly sampled points from the image. The saliency map is then treated as a binary classifier to separate the positive samples from the negative ones. Thresholding over the saliency map and plotting true positive rate vs. false positive rate results in the ROC curve.

AUC considers the saliency results at the locations of the human fixations. Thus, it distinguishes purely between a peaky saliency map and a smooth one. To view the fixations as samples of a distribution, rather than considering each fixation separately, similarly to [34], we prefer another metric: χ^2 distance between two distributions. The χ^2 distance prefers a peaky map over a smooth one.

For the χ^2, perfect prediction corresponds to a score of 0. For AUC, perfect prediction corresponds to a score of 1, while a score of 0.5 indicates the chance level. Thus, to be consistent, we use 1 - χ^2 when reporting our results.

To the best of our knowledge, we are the first exploiting depth for video saliency when viewing on 2D screens. Therefore, for a fair evaluation we compare our approach to the extended baseline algorithm (Sec. 3.1). We also compare our approach to video saliency technique [34], four image saliency methods [9, 18, 39, 45], depth-aware image saliency (RGBD) [32] and a Gaussian in the center [19].

Table 1 demonstrates a quantitative comparison using two different metrics: χ^2 and AUC. The "Ground-truth" in the bottom row is the upper bound for the saliency prediction, which measures how much the ground-truth fixation map "explains itself" (Equation 1). We use 38 out of 54 RGBD videos in our dataset for training, while the other 16 videos form the testing set. To quantify the impact of depth, we also carried out an experiment where we removed the depth information from the input to our approach.

The results of our depth-aware methods are the closest to the ground-truth. According to both χ^2 and AUC measures, the relative improvement over the state-of-the-art method [34] is approximately 15% (0.70/0.61). We also see that employing depth in video saliency algorithms, which are based on learning, improves their accuracy; both the extended baseline (Sec. 3.1) and our approach outperform previous approaches. Finally, the standard deviation of our approach is lower than in all other methods, making it more reliable than the others.
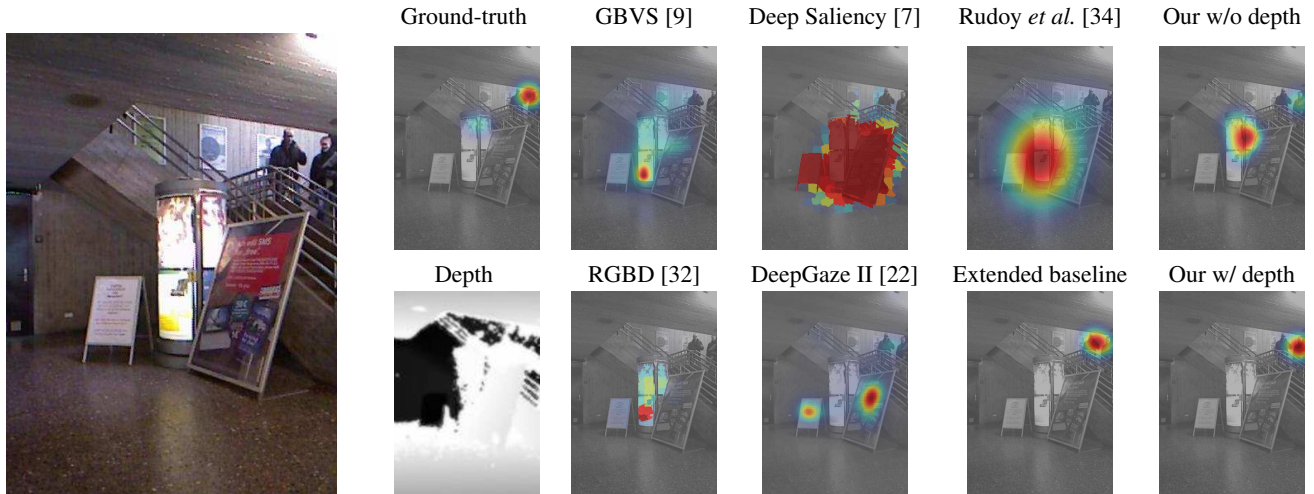
Method	$1 - \chi^2$	AUC
RGBD [32]	0.53 ± 0.21	0.66 ± 0.19
GBVS [9]	0.53 ± 0.25	0.65 ± 0.21
Judd *et al.* [18]	0.56 ± 0.21	0.67 ± 0.22
Vig *et al.* [39]	0.55 ± 0.23	0.64 ± 0.29
Zhang *et al.* [45]	0.57 ± 0.21	0.66 ± 0.25
Center [19]	0.56 ± 0.39	0.66 ± 0.36
Rudoy *et al.* [34]	0.61 ± 0.26	0.68 ± 0.23
Our approach (w/o depth)	0.60 ± 0.23	0.68 ± 0.21
Extended baseline	0.64 ± 0.22	0.70 ± 0.18
Our approach (w/ depth)	$\mathbf{0.70 \pm 0.15}$	$\mathbf{0.75 \pm 0.14}$
Ground-truth	0.84 ± 0.05	0.88 ± 0.06

Table 1. **Quantitative Evaluation.** We compare our method to depth-aware image saliency [32], four image saliency methods [9, 18, 39, 45], a Gaussian placed in the center [19], video saliency [34] and the extended baseline algorithm from Section 3.1. The upper bound (Ground-truth) for the saliency prediction is given in Equation 1. According to both χ^2 and AUC measures our method is the closest to the ground-truth, outperforming other state-of-the-art methods. Moreover, employing depth in learning based video saliency algorithms improves their accuracy.
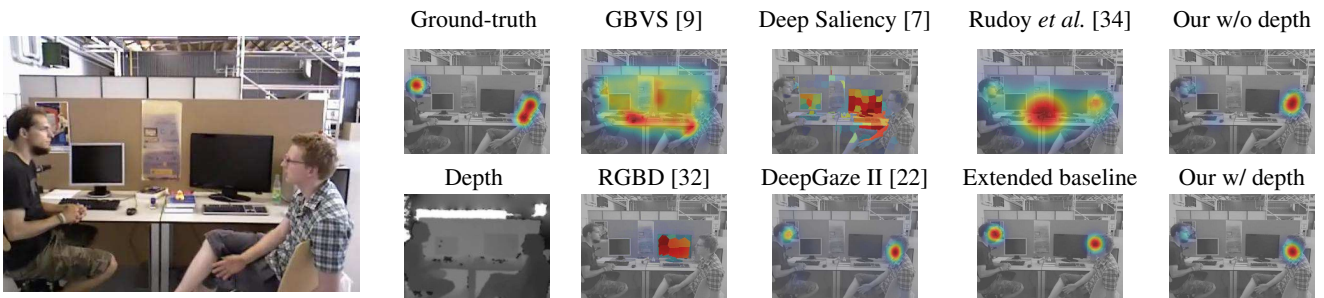
Note that the trivial approach of a Gaussian, placed in the center of the frame, produces fairly good average results due to two facts. First, when filming the videos we usually attempt to place the most interesting composition in the center of the frame. Second, when viewing relatively boring scenes we tend to move the gaze to the center of the frame. Thus, when comparing this trivial approach to the ground-truth we see relatively a high score in average. However, the standard deviation of this score is almost twice as high as the standard deviation of other methods, which makes the center-based Gaussian approach highly unreliable.

Qualitative evaluation: Figure 7 demonstrates a qualitative comparison of our approach to the ground-truth and other saliency techniques. In both cases the depth-aware saliency map is more visually consistent with the ground-truth than the maps of the other methods. For example, while watching a conversation between two persons, the gaze shifts from one face to the other, which is accurately captured by depth-aware saliency. The extended baseline is improved since the moving people in the corner are "salient enough", even in the low quality depth map. We refer the reader to the supplementary video, since motion of the people cannot be seen in the images.

In addition to the previously used saliency methods we compare our approach to DeepGaze II [22] and to Deep Saliency [7]. Unfortunately, the available implementation of these methods is too slow to run them on our entire dataset. As it can be seen, our results predict the ground truth better than these methods.

| Ground-truth | GBVS [9] | Deep Saliency [7] | Rudoy *et al.* [34] | Our w/o depth |
| Depth | RGBD [32] | DeepGaze II [22] | Extended baseline | Our w/ depth |

Depth-aware saliency maps are similar to the ground-truth, detecting background motion.

| Ground-truth | GBVS [9] | Deep Saliency [7] | Rudoy *et al.* [34] | Our w/o depth |
| Depth | RGBD [32] | DeepGaze II [22] | Extended baseline | Our w/ depth |

The gaze shifts between faces, which is accurately captured by depth-aware saliency.

Figure 7. **Qualitative evaluation on our dataset.** We compare our results to the ground-truth and to the additional saliency methods: video saliency (Rudoy *et al.*) [34], three image saliency methods (GBVS [9], DeepGaze II [22], Deep Saliency [7]) and depth-aware image saliency (RGBD) [32]. The left side of the figure demonstrates the input RGB frame, the depth data and the ground-truth, while the saliency results are shown on the right side. Both depth-aware methods outperform other state-of-the art methods. Our novel approach produces more concise results, and this fact is supported by the low standard deviation in Table 1.

6. Conclusion

In this paper, we proposed a novel depth-aware video saliency method, which predicts human foci of attention when viewing 3D video content on 2D screens. Our method employs a generative convolutional neural network to reconstruct saliency for each frame by implicitly learning the gaze transition from the previous frame. The network was trained to predict the saliency of the next frame by learning from depth, color, motion and saliency of the current frame. Experimental results show that exploiting depth is beneficial for video saliency, allowing our method to outperform previously proposed state-of-the-art methods. We believe that the significant boost with regard to the baseline is due to our net's ability to learn the nature of the depth maps. In this manner we capture the important parts of the depth maps that trigger gaze transitions.

Moreover, we presented Depth-Aware Video Saliency dataset, a comprehensive dataset of eye-fixation ground-truth for RGBD videos. This dataset contains videos representing common scenarios where depth-aware saliency is beneficial. To record eye-fixation ground-truth, we conducted a comprehensive user study, where the RGBD videos were displayed on regular screens ignoring depth information.

We believe that the constructed dataset and our work are helpful to stimulate further research in the area. In the future we plan to test our methods in various applications, *e.g.* video editing, video compression and video summarization.

Acknowledgements

References

[1] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Spatio-temporal convolutional sparse auto-encoder for sequence classification. In *BMVC*, 2012.

[2] A. Borji and L. Itti. Exploiting local and global patch rarities for saliency detection. In *CVPR*, 2012.

[3] A. Ciptadi, T. Hermans, and J. M. Rehg. An in depth view of saliency. In *BMVC*, 2013.

[4] X. Cui, Q. Liu, and D. Metaxas. Temporal spectral residual: fast motion saliency detection. In *ACM int. conf. on Multimedia*, 2009.

[5] K. Desingh, M. Krishna, D. Rajan, and C. Jawahar. Depth really matters: Improving visual salient region detection with depth. In *BMVC*, 2013.

[6] A. Dosovitskiy, J. T. Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. In *CVPR*, 2015.

[7] L. Gayoung, T. Yu-Wing, and K. Junmo. Deep saliency with encoded low level distance map and high level features. In *CVPR*, 2016.

[8] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *CVPR*, 2008.

[9] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in neural inf. proc. systems*, 2007.

[10] E. Herbst, X. Ren, and D. Fox. RGB-D flow: Dense 3-D motion estimation using color and depth. In *ICRA*, 2013.

[11] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[12] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. *Advances in neural information processing systems*, 21:681–688, 2008.

[13] Q. Huynh-Thu and L. Schiatti. Examination of 3D visual attention in stereoscopic video content. In *IS&T/SPIE Electronic Imaging*, 2011.

[14] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *Image Processing, IEEE Transactions on*, 13(10), 2004.

[15] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 20:1254–1259, 1998.

[16] L. Jansen, S. Onat, and P. König. Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, 9(1):29, 2009.

[17] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu. Depth saliency based on anisotropic center-surround difference. In *ICIP*, 2014.

[18] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009.

[19] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2010.

[20] K. Karsch, C. Liu, and S. Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *TPAMI*, 36(11):2144–2158, 2014.

[21] H. Kim, Y. Kim, J. Y. Sim, and C. S. Kim. Spatiotemporal saliency detection for video sequences based on random walk with restart. *IEEE Transactions on Image Processing*, 24(8):2552–2564, 2015.

[22] M. Kümmerer, T. S. A. Wallis, and M. Bethge. Deepgaze II: reading fixations from deep features trained on object recognition. *arXiv preprint*, arXiv/1610.01563, 2016.

[23] K. Lai, L. Bo, and D. Fox. Unsupervised feature learning for 3d scene labeling. In *ICRA*, 2014.

[24] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan. Depth matters: Influence of depth cues on visual saliency. In *ECCV*, 2012.

[25] Y. Liu, L. K. Cormack, and A. C. Bovik. Natural scene statistics at stereo fixations. In *Symposium on Eye-Tracking Research&Applications*. ACM, 2010.

[26] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen. Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1, 2017.

[27] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *TPAMI*, 32(1):171–177, 2010.

[28] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *ICANN*, 2011.

[29] P. Mital, T. Smith, R. Hill, and J. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, 2011.

[30] Y. Niu, Y. Geng, X. Li, and F. Liu. Leveraging stereopsis for saliency analysis. In *CVPR*, 2012.

[31] N. Ouerhani and H. Hugli. Computing visual attention from scene depth. In *ICPR*, 2000.

[32] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji. RGBD salient object detection: A benchmark and algorithms. In *ECCV*, 2014.

[33] C. Richardt, C. Stoll, N. Dodgson, H. Seidel, and C. Theobalt. Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos. *Computer Graphics Forum*, 2012.

[34] D. Rudoy, D. Goldman, E. Shechtman, and L. Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *CVPR*, 2013.

[35] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *AISTATS*, 2009.

[36] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12), 2009.

[37] N. Silberman, D. Hoiem, D. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012.

[38] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D slam systems. In *IROS*, 2012.

[39] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *CVPR*, 2014.

[40] J. Wang, P. Da-Silva, P. Le-Callet, and V. Ricordel. A computational model of stereoscopic 3D visual saliency. *Transactions on Image Processing*, 22(6):2151–2165, 2013.

[41] J. Wang, P. Le Callet, S. Tourancheau, V. Ricordel, M. P. Da Silva, et al. Study of depth bias of observers in free viewing of still stereoscopic synthetic stimuli. *Journal of Eye Movement Research*, 5(5), 2012.

[42] L. Wang, H. Lu, X. Ruan, and M. H. Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, 2015.

[43] S. Winkler and S. Ramanathan. Overview of eye tracking datasets. In *QoMEX*, 2013.

[44] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using SFM and object labels. In *ICCV*, 2013.

[45] J. Zhang and S. Sclaroff. Saliency detection: A boolean map approach. In *ICCV*, 2013.

[46] Q. Zhang, Y. Wang, and B. Li. Unsupervised video analysis based on a spatiotemporal saliency detector. *arXiv preprint arXiv:1503.06917*, 2015.

[47] S. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren. Video saliency detection via dynamic consistent spatio-temporal attention modelling. In *AAAI*, 2013.

[48] F. Zhou, S. Kang, and M. Cohen. Time-mapping using space-time saliency. In *CVPR*, 2014.

[49] Q. Zhou and V. Koltun. Dense scene reconstruction with points of interest. *TOG*, 2013.