Learning-based saliency model with depth information

Chih-Yao Ma

Department of Electronics Engineering, National Chiao-Tung University, Hsinchu, Taiwan
Present address: School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Hsueh-Ming Hang

Department of Electronics Engineering, National Chiao-Tung University, Hsinchu, Taiwan

Most previous studies on visual saliency focused on two-dimensional (2D) scenes. Due to the rapidly growing three-dimensional (3D) video applications, it is very desirable to know how depth information affects human visual attention. In this study, we first conducted eye-fixation experiments on 3D images. Our fixation data set comprises 475 3D images and 16 subjects. We used a Tobii TX300 eye tracker (Tobii, Stockholm, Sweden) to track the eye movement of each subject. In addition, this database contains 475 computed depth maps. Due to the scarcity of public-domain 3D fixation data, this data set should be useful to the 3D visual attention research community. Then, a learning-based visual attention model was designed to predict human attention. In addition to the popular 2D features, we included the depth map and its derived features. The results indicate that the extra depth information can enhance the saliency estimation accuracy specifically for close-up objects hidden in a complex-texture background. In addition, we examined the effectiveness of various low-, mid-, and high-level features on saliency prediction. Compared with both 2D and 3D state-of-the-art saliency estimation models, our methods show better performance on the 3D test images. The eye-tracking database and the MATLAB source codes for the proposed saliency model and evaluation methods are available on our website.

Introduction

Since the stereoscope was first invented, three-dimensional (3D) media has been recognized as an important next-generation visual media. Because the stereoscopic content can provide additional depth information and improve the viewing experience, many stereoscopic techniques ranging from 3D content acquisition to 3D display have been investigated. Although there has been a rapid growth in 3D research, 3D media still has a number of unsolved issues, and many of them are closely related to the human visual system (Yarbus, 1967; Posner, 1980).

One of the most noticeable research directions is 3D visual attention. For many applications in image processing, such as image cropping, thumbnailing, image search, quality assessment, and image compression, it is very useful to understand where humans look in a scene (Li & Itti, 2008; Judd, Ehinger, Durand, & Torralba, 2009). Our goal in this study was to construct a computational model for 3D attention. We first discuss the impact of binocular depth cues on human visual attention. Then, several existing 3D eye-tracking data sets and 3D visual attention models are briefly reviewed. One important part of our work is to design and collect the NCTU-3DFixation data set using stereo images and eye-tracking devices. To examine the consistency of our data set, we present an analysis on our data. Our main contribution is proposing a learning-based 3D saliency model and evaluating its performance on two data sets. In addition, the role of individual features and their combinations in our model is examined and reported.

3D visual attention

Although people are interested in attention modeling on stereoscopic 3D content, only a very small number of studies on this subject have been reported.

Citation: Ma, C.-Y., & Hang, H.-M. (2015). Learning-based saliency model with depth information. *Journal of Vision, 15*(6):19, 1–22, http://www.journalofvision.org/content/15/6/19, doi:10.1167/15.6.19.

doi: 10.1167/15.6.19 Received May 20, 2014; published May 26, 2015 ISSN 1534-7362

Influence of binocular depth cues on visual attention

In addition to the monocular cues that can induce depth perception, the stereoscopic videos provide binocular cues, enhancing our depth perception. Several studies were reported on how human attention may be affected by binocular depth cues. Jansen, Onat, and König (2009) investigated the influence of disparity on human attention based on their two-dimensional (2D) and 3D still-image experiments. Their results show that the additional depth information leads to an increased number of fixations, shorter and faster saccades, and increased spatial extents of exploration. Therefore, they concluded that depth information changes the basic eye-movement patterns and, thus, that the depth map can be an essential image salient feature.

By presenting the 2D and 3D versions of the same video content to viewers, Häkkinen, Kawai, Takatalo, Mitsuya, and Nyman (2010) demonstrated how stereoscopic content could affect eye-movement patterns. Their results suggest that the eye movements on 3D content are more widely distributed. Huynh-Thu and Schiatti (2011) also examined the differences in visual attention between 2D and 3D content. However, different from Jansen et al. (2009), Huynh-Thu and Schiatti (2011) used video clips rather than still images. The average saccade velocity was found to be higher when viewing 3D stereoscopic content, which is consistent with the results reported by Jansen et al. (2009).

Although it is generally agreed that images with higher luminance variation attract humans' attention, Liu, Cormack, and Bovik (2010) found that the higher variations in disparity gradient and contrast somehow create a forbidden zone, where the left-eye and right-eye images cannot be fused properly by the human brain. Therefore, human subjects do not pay attention to the areas with high variations in disparity. Several studies were cited in their paper to support their proposition.

The previous findings indicate that human eye-movement patterns are influenced by both the image or video content and the values of disparity. However, our experiments showed that the difference in watching 3D images (vs. 2D images) is most significant only for the first three fixations. Once the viewing time is sufficiently long, the 2D low-level features still dominate human visual attention.

3D fixation data set

The lack of a 3D fixation data set with ground truth has limited the development of a 3D visual attention model. To the best of our knowledge, so far only three published data sets contain 3D images, depth maps, and eye-fixation data as below.

Jansen data set

Jansen et al. (2009) compared the fixation difference on 2D and 3D still images. They recorded binocular eye-movement data on viewing the 2D and 3D versions of natural, pink-noise, and white-noise images. However, their data set contained only 28 images, which is insufficient to train a learning-based saliency model.

3DGaze

Wang, Da Silva, Le Callet, and Ricordel (2013) created and published an eye-tracking database containing 18 stereoscopic images, their associated disparity maps, and the eye-movement data for both eyes. The stereoscopic images in their database were acquired from two sources: (a) 10 images came from the Middlebury 2005/2006 image data set, and (b) eight images, which were captured by the authors, came from the IVC 3D image data set. Similar to the Jansen data set, this data set size is quite small.

NUS-3DSaliency

Lang et al. (2012) described a fairly large human eye-fixation database, which contains 600 2D versus 3D image pairs viewed by 80 subjects. The depth information came directly from a Kinect depth camera, and the eye-tracking data were captured in both 2D and 3D free-viewing experiments. The range of the Kinect depth sensor is limited to about 4 m, and because the Kinect depth sensor is strongly affected by ambient lighting, it is not suitable for outdoor scenes. Therefore, the smoothed depth maps are inaccurate for some cases, and the variety of the scenes is somewhat restricted. Another issue is that all the 3D stimuli were generated by virtual view synthesis, which is sensitive to depth map errors. The authors tried to remove the artifacts in the depth maps and carefully picked up the better quality images. Thus, the data set should be quite useful; however, the synthesized 3D images of this data set are not available to the public.

Therefore, we conducted 3D image fixation experiments of our own and compiled a set of data. We posted these data on the Internet, which hopefully will be useful to the researchers studying this subject.

3D visual attention models

Currently, only a few computational models of 3D visual attention can be found in the literature. Nearly

all of these models contain a 2D stage in which the 2D visual features are extracted and used to compute the 2D saliency maps. This structure matches the findings in Jansen et al. (2009). They found that there is no significant difference between the viewing of 2D and 3D stimuli regarding the 2D saliency map derived based on 2D visual features. This consistence of 2D low-level visual features for 2D and 3D stimuli implies the possibility of adapting the existing 2D visual attention models to the 3D visual attention model.Therefore, Wang et al. (2013) classified these 3D attention models into three categories depending on how they used the depth information.

Depth-weighting models

Apart from detecting the salient areas by using the 2D visual features, these models add a step in which the depth information is used as a multiplicative weighting factor in generating the 3D saliency map (Maki, Nordlund, & Eklundh, 1996; Chamaret, Godeffroy, Lopez, & Le Meur, 2010; Y. Zhang, Jiang, Yu, & Chen, 2010; Lang et al., 2012). Lang et al. (2012) analyzed the major discrepancies between the 2D and 3D human fixation data of the same scenes, which are further abstracted and modeled as novel depth priors. In order to determine saliency, Lang et al. extended seven existing models to include the learned depth priors. The final saliency can be achieved by simply using summation or point-wise multiplication as the fusion operation of two components. By using different evaluation methods, they observed a 6% to 7% increase in prediction accuracy using depth priors.

Y. Zhang et al. (2010) proposed a bottom-up stereoscopic visual attention model to simulate the human visual system. Spatial and motion saliency maps were constructed from features such as color, orientation, and motion contrasts. Then, a depth-based dynamic fusion was used to integrate these features. However, only the attention detection experiment was performed to evaluate the performance of their model. There are no actual eye-tracking data to determine the accuracy of their model.

Depth-saliency models

This type of model first extracts depth-saliency maps based on the depth map. These depth-saliency maps are then combined with the 2D saliency maps using a pooling strategy to produce a final 3D saliency map (Ouerhani & Hugli, 2000; Potapova, Zillich, & Vincze, 2011; Wang et al., 2013).

Wang et al. (2013) proposed a model that takes depth as an additional visual feature. The measure of depth saliency is derived from the eye-movement data obtained from eye-tracking experiment using synthetic stimuli (Wang, Le Callet, Tourancheau, Ricordel, & Da Silva, 2012). Wang et al. (2013) believed that there are several advantages of using the synthetic stimuli. First, it can precisely control the depth of the object and background. Second, the influence of 2D visual features on viewing behavior and the influence of monocular depth cues can be restricted. Instead of directly using the depth map, the authors used a probability-learning algorithm to model the relationship between the depth contrast (applying a Difference of Gaussians (DoG) filter on the depth map) of each position and the probability of this position being gazed at. Finally, by combining the depth-saliency map and the 2D saliency map, the predicted saliency map was generated. Their results show that the depth information can actually improve the performance of using the 2D saliency model only.

Stereovision models

This type of model takes into account the mechanism of stereoscopic perception in the human visual system (HVS). Bruce and Tsotsos (2005) extended the existing 2D attention architecture to the binocular domain, in conjunction with the connectivity of units involved in achieving stereovision, but no detailed implementation or evaluations were reported.

Kim, Sanghoon, and Bovik (2014) combined the well-known perceptual and attentional principles with the traditional bottom-up low-level features. Then, they came up with a detailed and sophisticated saliency prediction model for stereoscopic videos. This model can be considered as a combination of the three types of 3D visual attention models mentioned above. It first produces 3D space–time salient segments (regions) in a video sequence and then calculates the saliency strength of different scene types (classified based on motion information). More importantly, the authors used two additional visual factors (foveation and Panum's fusional area) to increase the prediction precision.

Most of the existing 3D visual attention models belong to the first (depth-weighting model) and second (depth-saliency model) categories. Although the depth-weighting model can easily adopt the existing 2D models, it may miss certain salient regions signified by the depth features only. On the other hand, the depth value alone is not a reliable attention cue. In a number of scenes the floor has a smaller depth value because it is close to the camera, but the observers are often not interested in the floor. Therefore, combining the depth map directly with the 2D saliency map sometimes misidentifies the salient area.

Although several computational models of 3D visual attention have been proposed, most of these works did

not report subjective experimental results in evaluating the proposed models (Maki et al., 1996; Bruce & Tsotsos, 2005; Chamaret et al., 2010; Potapova et al., 2011; Wang et al., 2012). The very first challenge of modeling the 3D visual attention is how to reliably collect and interpret eye-tracking data. Most studies on 3D visual attention have used the tracking data recorded from only one eye. Wang et al. (2013) argued that binocular recordings are necessary for recording 3D gaze despite the fact that such eye-tracking equipment can provide only a 2D spatial gaze location individually for each eye.

One plausible approach is using two images of the same scene obtained from slightly different viewing angles. It is then possible to triangulate the distance to an object with a high degree of accuracy. However, using the triangulation of two 2D gaze points from both eyes to produce a single 3D gaze point is highly dependent on the calibration accuracy of the system. In the case of an experiment using 3D stimuli, it is difficult to ensure that the observer accurately looked at the point at the given depth plane. Thus, further studies are needed to specify the standard protocols for conducting eye-tracking experiments, and a reliable procedure is needed for analyzing the eye-tracking data.

Eye-fixation database

Our NCTU-3DFixation data set comprises 475 3D images along with their depth maps and the eye-fixation data. The 3D images, eye-tracking data, and accompanying codes in MATLAB are all available at http://cwww.ee.nctu.edu.tw/wiki/core/pmwiki.php?n=People.HangResearchMaterial.

Image content

Instead of using the virtual view synthesis techniques to generate the test images (stimuli), or dataset aims to provide the same viewing experience when the users watch regular entertainment 3D videos. Therefore, our 3D content mainly came from existing 3D movies or videos. Figure 1 shows the 11 sequences collected, including four movie trailers, three 3D videos from 3dtv, and four videos from YouTube. We captured the image frames randomly from the left- and right-eye videos separately and interlaced a 3D image pair into a full high definition (HD) image row by row. We carefully sieved out the distorted or unnatural images and selected 475 good-quality 3D images as our test image database.

Depth information

One of the most challenging parts of constructing a 3D eye-fixation database is the depth information, and this may explain why only a few eye-fixation databases for 3D images exist. The depth maps in the NUS-3DSaliency data set came from the Kinect depth sensor. As discussed earlier, it imposes some limitations. Because our collected 3D images have only pictures (no depth maps), our depth maps are generated using Depth Estimation Reference Software (DERS; version 5.0) provided by the International Telecommunication Union/Moving Picture Experts Group (ITU/MPEG) standard committee (Tanimoto, 2012). Some sample depth maps together with their pictures are shown in Figure 1.

To achieve a higher accuracy, the original DERS software uses three camera views to generate one depth map for the center view. However, in our case, we have images only from the left and right views. Therefore, we treat the right image as the center view used by the software, and the left image is fed to both the left and right views in the software. By doing so, the produced depth map is located on the viewing point of the right image. (It also requires some additional modifications to the DERS source code.) We choose the right view as the base (or reference) because approximately two-thirds of the population is right-eye dominant. That is, our sense of spatial structure is more likely based on the viewpoint of the right eye. (Because the same left and right images are used to estimate the depth maps, the right-view depth map and the left-view depth map contain the same amount of information.) This is different from many other visual experiments, such as the work by Wang et al. (2013) and many others, which choose the left image as the base image.

Data collection

In constructing our data set, we used a Tobii TX300 eye tracker (Tobii, Stockholm, Sweden) to accurately track the eye movements of the subjects (see Figure 2). Note that the sampling rate of the Tobii TX300 is up to 300 Hz, but in our experiment we set it at 120 Hz due to the use of polarization glasses, which seem to decrease the tracking reliability at higher sampling rate. This system allows the slight movement of viewers' heads.

We recorded the eye-tracking data of 16 users (subjects). All 3D images were displayed on a 23-in. patterned retarder 3D display (Asus VG236H, Asus, Taipei, Taiwan) at a resolution of 1920×1080, a refresh rate of 120 Hz, a brightness of 400 cd/m^2, and a contrast ratio of 100,000:1. The average viewing distance was 78.5 cm. The interocular crosstalk was

Figure 1. The 475 3D images captured from 11 3D videos, which come from YouTube and 3dtv. The depth maps are generated by DERS. The figure shows the original images (first row), the generated depth maps (second row), and the human fixation density maps (third row) in the NCTU-3DFixation database.

imperceptible (<1%) when watching directly in front of the center of the display.

All subjects (both males and females; aged 18–24 years) were asked to precisely complete the nine-point calibration so that the eye tracker could flawlessly locate the gaze point at nearly any location on the screen. Only the subjects whose tracking error was less than 5 mm (about 0.37°) were included in our database.

Figure 2. Experimental setup: 3D images are displayed on a 23-in. 3D display, and the Tobii TX300 eye tracker is used to track the user's eye movement at a 120-Hz sampling rate.

The eye tracker successfully tracked eye movement, with an average of 95.7% of experiment time for all 16 subjects. The lowest tracking time was 90.1% of experiment time and the highest was 99.6%. All the experiments were conducted in a dark room to provide the best sense of depth for 3D images and to reduce the distraction caused by other objects in the room.

Using the E-Prime 2.0 software (Psychology Software Tools, Pittsburgh, PA), the 475 chosen 3D images were randomly divided into five sections. Every picture was displayed for 4 s, and a transition picture with a fixed white cross on the center was displayed for 2 s as illustrated in Figure 3. Users were asked to take a 10-min break between two viewing sessions to reduce fatigue. Our presentation time for every stimulus was set to 4 s, which is similar to several other experiments reported: 6 s for the NUS-3DSaliency data set (Lang et al., 2012), 2 s for the FIFA data set (Cerf, Frady, & Koch, 2009), 4 s for the Toronto database (Bruce & Tsotsos, 2009), 3 s for the Massachusetts Institute of Technology (MIT) database (Judd et al., 2009), and 5 s for the NUSEF database (Ramanathan, Katti, Sebe, Kankanhalli, & Chua, 2010). The only exception is the 3DGaze eye-tracking database (Wang et al., 2013), in which the presentation time is set to 15 s, which is much longer than the others. Additionally, the literature shows that compared with the 2D experiments, the additional depth sensation leads to an increased number of fixations, shorter and faster saccades, and an increased spatial extent of exploration (Jansen et al., 2009; Häkkinen et al., 2010). Therefore, we adopted the 4-s presentation time, which seems to be adequate for 3D visual attention research (further discussed later).

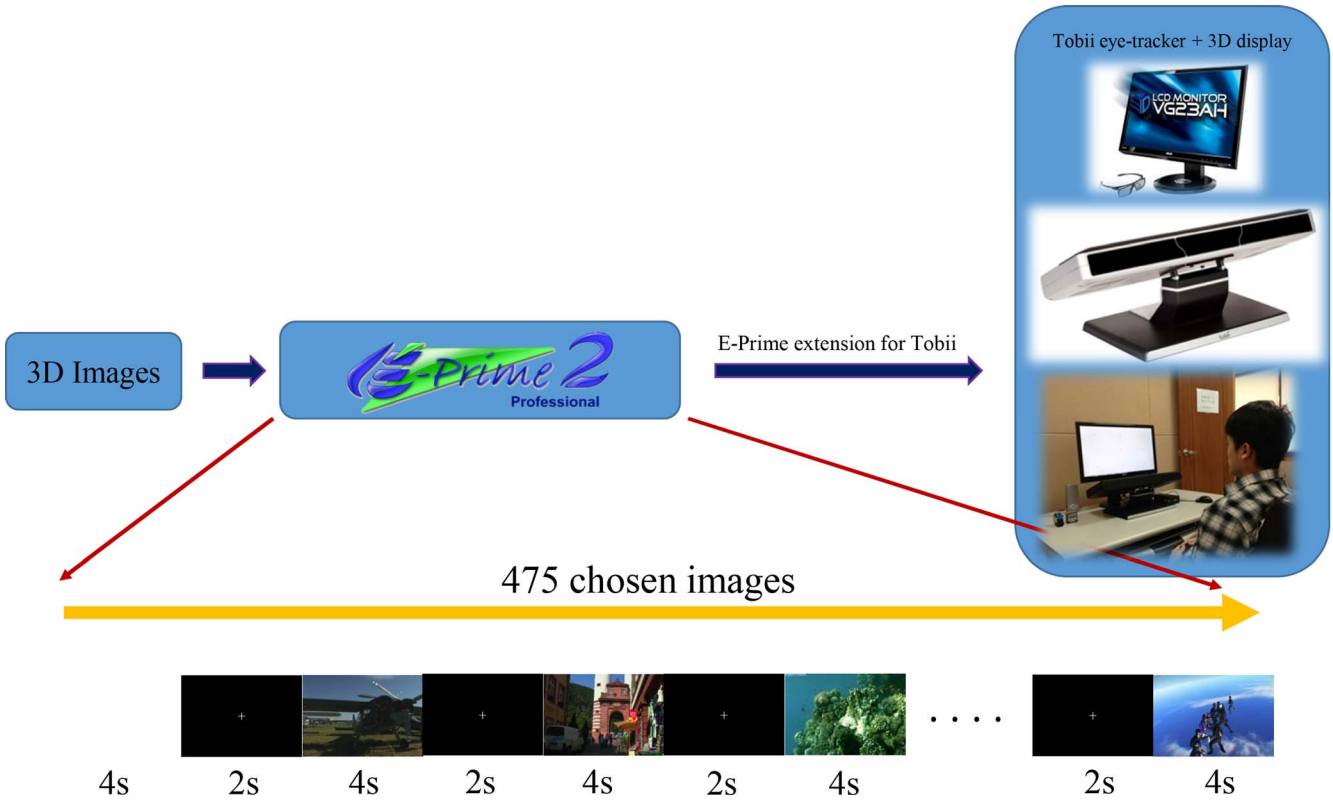

Figure 3. Testing procedure: The 475 chosen images are randomly divided into five sections by E-Prime. Each image is displayed sequentially, with a transition (dummy) picture between two presented images.

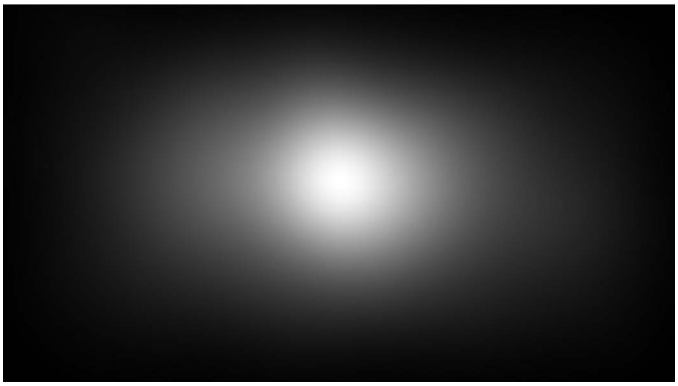

Figure 4. Average human density map over all 475 images (16 subjects).

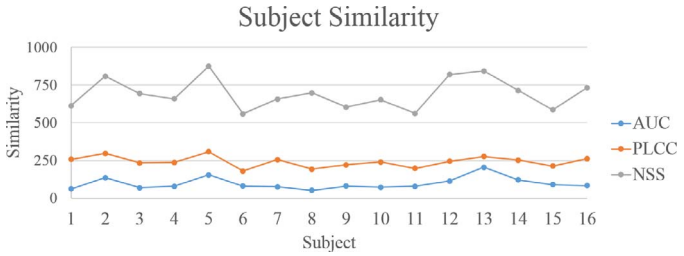

Figure 5. Similarity scores of AUC, PLCC, and NSS for individual subjects.

Human fixation density map

The fixation maps were constructed directly from the recorded fixation points of all subjects. The transition picture had only a white cross in the center, the purpose of which was to keep the user's attention on the center of the screen. Therefore, the first detected fixation point at the beginning of each image presentation was the screen center, which was image content independent.

In order to produce a continuous fixation map of an image, we adopted a smoothing technique that convolves a Gaussian filter with all fixation points. In a way, the recorded fixation points are samples of the Gaussian distributions on the ground truth fixation map. Examples of fixation density map (with Gaussian smoothing) for 11 images are shown in Figure 1. The brighter pixels denote the higher fixation values (higher probability). Because different values of the sigma parameter in the Gaussian smoothing filter produce different density maps, its value consequently affects the saliency model derived based on the fixation map. In order to simulate the human visual system, we set the sigma parameter of the Gaussian filter the same as the size of the fovea. According to our experimental setup, the sigma value was set to 96 for our fixation density maps.

Analysis of database

This database aims at quantitative analysis of fixation points and gaze paths and provides the ground truth data for saliency model research. Therefore, we examined the reliability and consistency of the collected data. We also estimated its upper theoretical performance limit (UTPL; Stankiewicz, Anderson, & Moore, 2011). We followed the procedure suggested by Stankiewicz et al. (2011). The MATLAB codes for this evaluation can be found at http://cwww.ee.nctu.edu.tw/wiki/core/pmwiki.php?n=People.HangResearchMaterial.

Center bias

In our data set we observed a strong center bias, which previously has been reported by the other eye-tracking data sets (Tatler, 2007; L. Zhang, Tong, Marks, Shan, & Cottrell, 2008). Figure 4 shows the average human density map from all 475 images. This center bias phenomenon is often attributed to the setup of the experiments, in which subjects sit at a central location in front of the screen and the objects of interest tend to be placed in the center of an image frame. We also noticed that a center bias component is essential for any saliency model, as discussed in the works of Judd et al. (2009) and Zhao and Koch (2011).

Individual subject consistency

We adopted several metrics, including the area under the receiver operating characteristic (ROC) curve (AUC), Pearson linear correlation coefficient (PLCC), and normalized scanpath salience (NSS), to evaluate the difference between two subjects (Peters, Iyer, Itti, & Koch, 2005; Stankiewicz et al., 2011). By definition, NSS evaluates the (predicted) salience values at the fixation locations (Peters et al., 2005). It extracts all the predicted saliency values at the observed fixation locations along a subject's scanpath and then averages these values to calculate the NSS score.

To check the consistency of individual subject fixations, we used the fixation density map of one subject as the predictor to predict the fixation density maps of the other 15 subjects. The total similarity for one reference subject is computed as follows.

Variable	AUC	PLCC	NSS	Similarity	EMD
UTPL	0.893	0.928	2.537	0.803	1.463

Table 1. The UTPL of the NCTU-3DFixation data set. The results were calculated by five metrics: AUC, PLCC, NSS, similarity, and EMD.

$$s_{k,n}: \text{Matching score of the } n\text{th user to predict the others on the } k\text{th image}$$

$$S_{k,n} = [\, s_{k,1} \quad s_{k,2} \quad \cdots \quad s_{k,16} \,]$$

$$similarity_{k,n} = \begin{cases} \max(S_{k,n}) - \min(S_{k,n}), & \text{when } \min(S_{k,n}) = s_{k,n} \\ 0, & \text{when } \min(S_{k,n}) \neq s_{k,n} \end{cases}$$

$$Similarity_n = \sum_{k=1}^{475} similarity_{k,n}$$

Thus, $S_{k,n}$ is the score of how well the nth user predicts the others (on the kth image). The $similarity_{k,n}$ for each subject is set to the difference between the maximum and the minimum scores, if the score of the nth subject is the minimum of all the subjects; otherwise, it is assigned to zero. Each similarity score of one subject for AUC, PLCC, and NSS is the sum of all 475 images. As shown in Figure 5, a lower similarity represents that the subject's data are rather different from that of the others, and a higher score indicates that this subject agrees with the others in the attention regions. This similarity score helps in judging the data reliability of a subject.

Upper theoretical performance limit

We also computed the UTPL, which indicates the consistency of the data (Stankiewicz et al., 2011). The results are listed in Table 1. The UTPL is the similarity between the fixation density map obtained from half of the human observers (randomly selected) and the fixation density map produced by the other half of the observers. If the data set is very consistent, the UTPL value should be close to the maximum of that similarity measure. If the data set itself is not consistent, the learning based saliency model derived from this set of data may not be accurate.

In calculating the UTPL, in addition to AUC, PLCC, and NSS, we used similarity and Earth's mover distance (EMD; Rubner, Tomasi, & Guibas, 2000; Pele & Werman, 2009; Judd, Durand, & Torralba, 2012). Similarity is a measure of how similar two distributions are. After each distribution is normalized, the similarity measure is the sum of the minimum values at each point in the distributions. EMD captures the global discrepancy of two distributions. That is, given two distributions, EMD measures the least amount of work needed to convert one distribution to the other one. We repeated this process 100 times for each test image to achieve a robust estimation. The higher score indicates a more consistent data set. Except for the EMD score, the scores of AUC, PLCC, and similarity range from 0 to 1. The results, which are shown in Table 1, indicate that our data set is rather consistent. (One may compare our results here with similar works of Engelke et al., 2013, and Wang et al., 2013.)

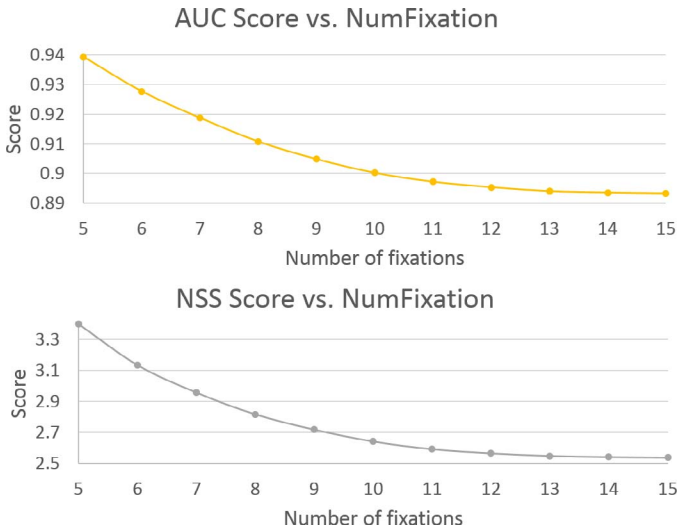

Figure 6. Scores of AUC and NSS versus the number of fixations.

Fixation numbers versus score

As discussed in the literature, the divergence of human fixation location increases when the presentation time in the subjective test increases. Therefore, we also examined the AUC and NSS values against the (time-ordered) number of fixations. As depicted in Figure 6, the consistency of human fixation data decreases as the time extends (i.e., more fixation points are included). Note that the initial fixation point (picture center) is ignored; thus, "the first five fixations" actually means the first six fixations without the first one. The average number of fixations of each stimulus for all subjects and images of our data set was 10.

Learning-based 3D saliency model

In the previous sections, we discussed the importance of depth information in 3D saliency modeling. Different from the previous 3D attention models that combine the existing 2D model directly with the depth map, our scheme finds the best weighting of the existing 2D features together with a possibly new type of 3D (depth) feature. We believe that integrating with the

extra depth information may change the original 2D saliency model.

For this purpose, we used a learning-based saliency model design similar to that of Judd et al. (2009). They collected the eye-tracking data of 15 viewers on 1,003 images and used this database as training and testing samples to produce a saliency model making use of the low-, mid-, and high-level image features. We extended their work by modifying, replacing, and adding new image features to their model and conducted training using the 3D images.

Features used in machine learning

Low-level features

All the features we use are discussed in this subsection. One sample image and its corresponding features are shown in Figure 7. For the low-level features, we adopted the local energy of the steerable pyramid filters, which has been found to be physiologically plausible and strongly correlated with visual attention (Simoncelli & Freeman, 1995). Additionally, we adopted the features that are combinations of all subband pyramids suggested by Rosenholtz (1999) and Oliva and Torralba (2001). Our feature set also included the three channels (intensity, color, and orientation) proposed by Itti and Koch (2000). We discovered that the color channels used by Judd et al. (2009) are extremely important for predicting saliency regions. These color channels include the values of the red, green, and blue channels as well as the probabilities of each of these channels and the probability of each color as computed from the 3D color histograms of the image filtered with a median filter at six different scales.

Mid-level features

According to the work by Judd et al. (2009), because most objects rest on the surface of the Earth, the horizon is a place humans naturally look for salient objects. Judd et al. (2009) thus train a detector from mid-level gist features to detect the horizon line (Oliva & Torralba, 2001). We also included this feature in our model.

High-level features

Many studies showed that viewers are more likely to fixate on faces and persons. Thus, we used the face detector proposed by Viola and Jones (2001) and the Felzenszwalb person detector (Felzenszwalb, McAllester, & Ramanan, 2008) to locate the region of the humans face and body. After these detectors determined a location, a Gaussian distribution was placed on that location to spread out the probability distribution of a salient region. The scale parameter of the distribution was set to the window size of the located region.

Center bias

A strong center bias, which has been previously reported, dominates the observational behavior in the existing eye-tracking data sets (Tatler, 2007; L. Zhang et al., 2008). We also observed a strong center bias from the average fixation density map of all 475 images in our database. Thus, we included a center bias feature in our 3D saliency model.

Depth features

According to the study of Wang et al. (2013), the same disparity value can produce different perceived depth due to the other viewing conditions. Therefore, they proposed a transformation that maps an original disparity map to a perceived depth map. Also, they introduced a depth-based saliency map, which is generated by applying a difference of Gaussian filter to the depth map based on the assumption that the depth is noticed mostly at surface discontinuities (Didyk, Ritschel, Eisemann, Myszkowski, & Seidel, 2011).

However, we found that viewers typically focused on the body of interesting objects rather than the boundaries of the objects (also reported by Liu et al., 2010). Therefore, in addition to the original depth map and the normalized depth map, we applied steerable pyramid filters to the depth map to increase the depth predictability. The idea is that the depth-related spatial structure can be represented by the orientation features at different scales. The combinations of these different-scale orientations give a clue in finding the locations of the corresponding salient objects. More details are described in the Discussion section.

In addition to the original depth, the normalized depth map, and the depth-based orientation information, we added one more set of depth features. We multiplied the depth features (different orientations and scales) by the normalized depth map pixel by pixel. This can be viewed as the combination of a depth-weighted model and a depth-directed saliency model.

CovSal features

A concept called region covariance was introduced by Erdem and Erdem (2013) to address the issue of how separate feature maps are combined to produce a master saliency map. In their model, so-called CovSal, only the Lab color feature, the orientation, and pixel location were used to represent an image pixel. They used the covariance matrices of simple image features as the metafeatures for saliency estimation. We found

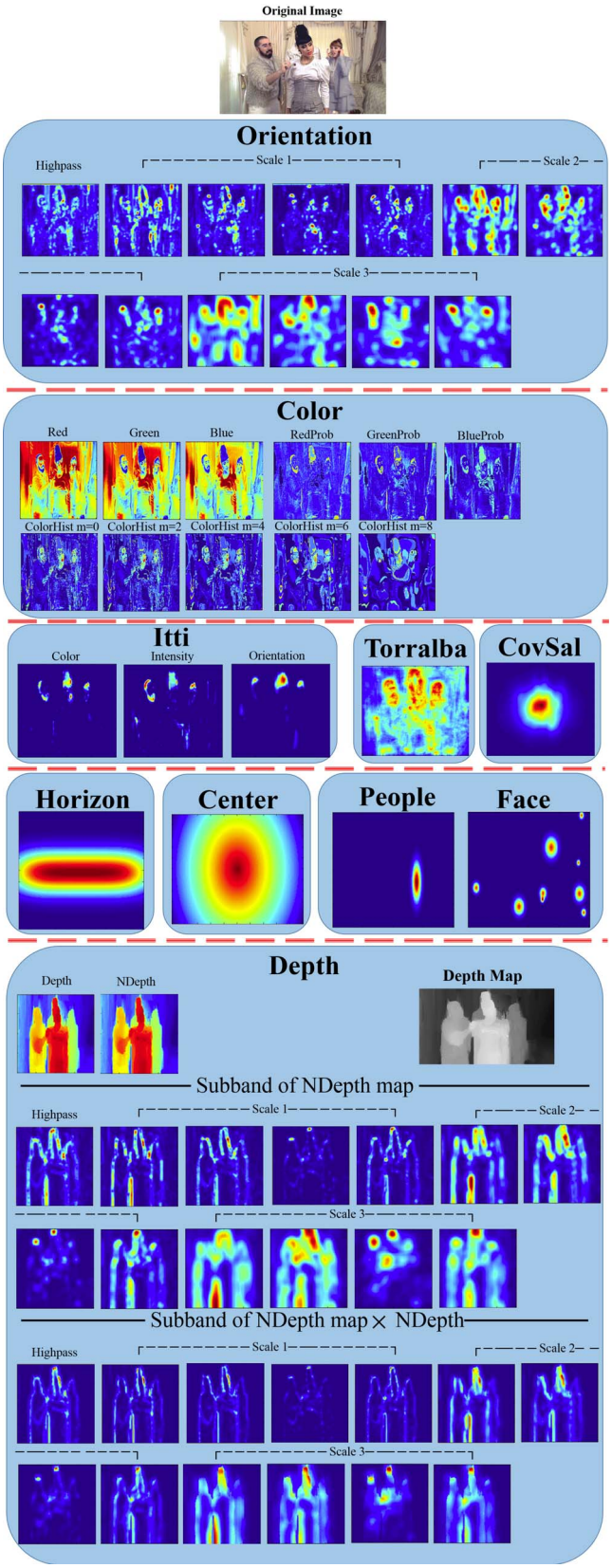

Figure 7. Features maps used in our model, including orientation, color, Itti channels, Torralba, CovSal, horizon, center, people, face, and depth information.

that their saliency estimation method has good potential to identify interesting objects of normal size (not too small or too large), especially when the object number is small. Thus, we include the output of the CovSal model as a feature (CovSal feature) to locate interesting objects.

All of the above features have different effects on the performance of the saliency modeling or estimation. These factors are further discussed in the Performance section.

Training by support vector machine

In order to train our saliency model, we adopted a support vector machine (SVM) classifier developed by Chang and Lin (2011) to predict the possibility of a pixel being fixated. The ground truth data came from our NCTU-3DFixation database described earlier.

We divided the 475 images into 425 training images and 50 testing images. On each image, we chose 20 pixels labeled as positive samples from the top 5% salient regions of the merged human fixation density map and 20 pixels labeled as negative samples from the bottom 70% salient regions. We found that the choices of the percentage of positive and negative samples are very important. Judd et al. (2009) chose the top 20% salient region for their positive samples because their ground truth distribution (human fixation density map) is less diverse. In our case, the choice of top 5% for the positive samples would result in more precise allocation of the saliency regions.

Every image was resized to 200×200 pixels before feature extraction, and every image feature was directly extracted from the right view of a stereo pair because the depth maps of the data set are generated for the right view as discussed before.

We applied the same normalization process as in Judd et al. (2009) to the feature space. In order to have zero mean and unit variance of each feature, we normalized each feature separately in the training data and applied the same normalization parameters to the test data.

In SVM algorithm selection, we chose the linear kernel over the radial basis function kernel because they both perform about equally well but the linear model requires less computational time than the radial basis function kernel. We used a simple greedy algorithm with the misclassification cost parameter set to 6.

Evaluation methods

Several evaluation methods for measuring the accuracy of a saliency model have been proposed. After

Model	AUC	PLCC	NSS	Similarity	EMD
Proposed model	0.837	0.688	1.594	0.562	2.430
Judd et al. (2009)	0.829	0.621	1.367	0.540	2.544
CovSal (Erdem & Erdem, 2013)	0.811	0.650	1.434	0.620	2.743
Proposed model without CovSal	0.837	0.629	1.400	0.531	2.555

Table 2. Performance comparison on our data set, including the proposed model, the model of Judd et al. (2009), and the CovSal model.

some investigation, we choose five evaluation metrics: AUC (Stankiewicz et al., 2011), PLCC and NSS (Peters et al., 2005), similarity (Judd et al., 2012), and EMD (Rubner et al., 2000; Pele & Werman, 2009).

Performance

The performance of our model was tested on two databases. First, the model was tested on our NCTU-3DFixation database and was compared with the other top saliency models. Second, we applied our model to the 3DGaze database created and published by Wang et al. (2013).

Performance on the NCTU-3DFixation database

In the experiments on the NCTU-3DFixation data set, the training and testing processes were repeated 20 times (20 experiments) to obtain robust results. In each experiment, 425 randomly selected images were used for training and the rest were testing images. We computed the evaluation metrics (scores) of the testing images for each experiment separately; the average scores of 20 experiments are shown in the first row of Table 2. The weighting parameters of our final model come from the average of the weighting parameters in these 20 experiments. For comparison, we also tested the proposed model without the CovSal feature as well as the models proposed by Judd et al. (2009) and Erdem and Erdem (2013) on the same data set. These two models are currently two top-rated 2D models on the saliency benchmarks (Judd et al., 2012). However, on the 3D data set, our model outperformed these two models for nearly all the commonly used performance indices. Some predicted saliency maps are shown in Figure 8. Both our model and the model of Judd et al. (2009) have the center bias feature; therefore, we added a center bias to the predicted saliency map of the CovSal model because their original center bias was multiplied by a Gaussian distribution with a presaliency map, and this multiplicative center bias sometimes decreases the performance of their model. As discussed earlier, the CovSal model is able to locate nearly all salient objects in an image, and indeed the CovSal feature can improve both PLCC and NSS metrics.

Note that, first, the car detection in the model of Judd et al. (2009) is discarded due to its long computing time and small contribution to the performance. Second, the models of Judd et al. (2009) and Erdem and Erdem (2013) were originally designed for 2D images (not 3D images). Thus, they may not perform as well as our model on the 3D data. On the other hand, the only extra information associated with the 3D images is the depth map. We next discuss that the depth information is important for certain cases but that, on average, excluding the depth information in our model does not reduce the performance metrics much.

2D model versus 3D model

In our experiments, we found that the major difference between watching 2D and 3D content typically appears in first three fixations. After the first three fixations, the distribution of the accumulated fixation density map converges to a stable map, as shown in Figure 9. Note that the (first) initial fixation point (i.e., picture center) is ignored, and thus "the first three fixations" actually means the first two fixations without the first one. Generally, the fixation map convergence is achieved in 4 s (image display time) and there are about 10 fixations for each image. During this 4-s period, human observers tend to pay attention to the most interesting objects in the first few fixations. Then their eyes move to the other areas in a test image and later focus on the most interesting objects again.

To further explore the 2D and 3D modeling differences, we compared the Judd model and ours against the number of fixations as shown in Figure 10. In this plot only NSS and EMD are displayed because, among five evaluation metrics, they show more significant differences between these two models. Indeed, our model (3D model) shows better performance on the first few meaningful fixations. With more fixations, the EMD metric favors the Judd model and the NSS metric favors ours. Overall, considering the other metrics as well, our model produces somewhat better results.

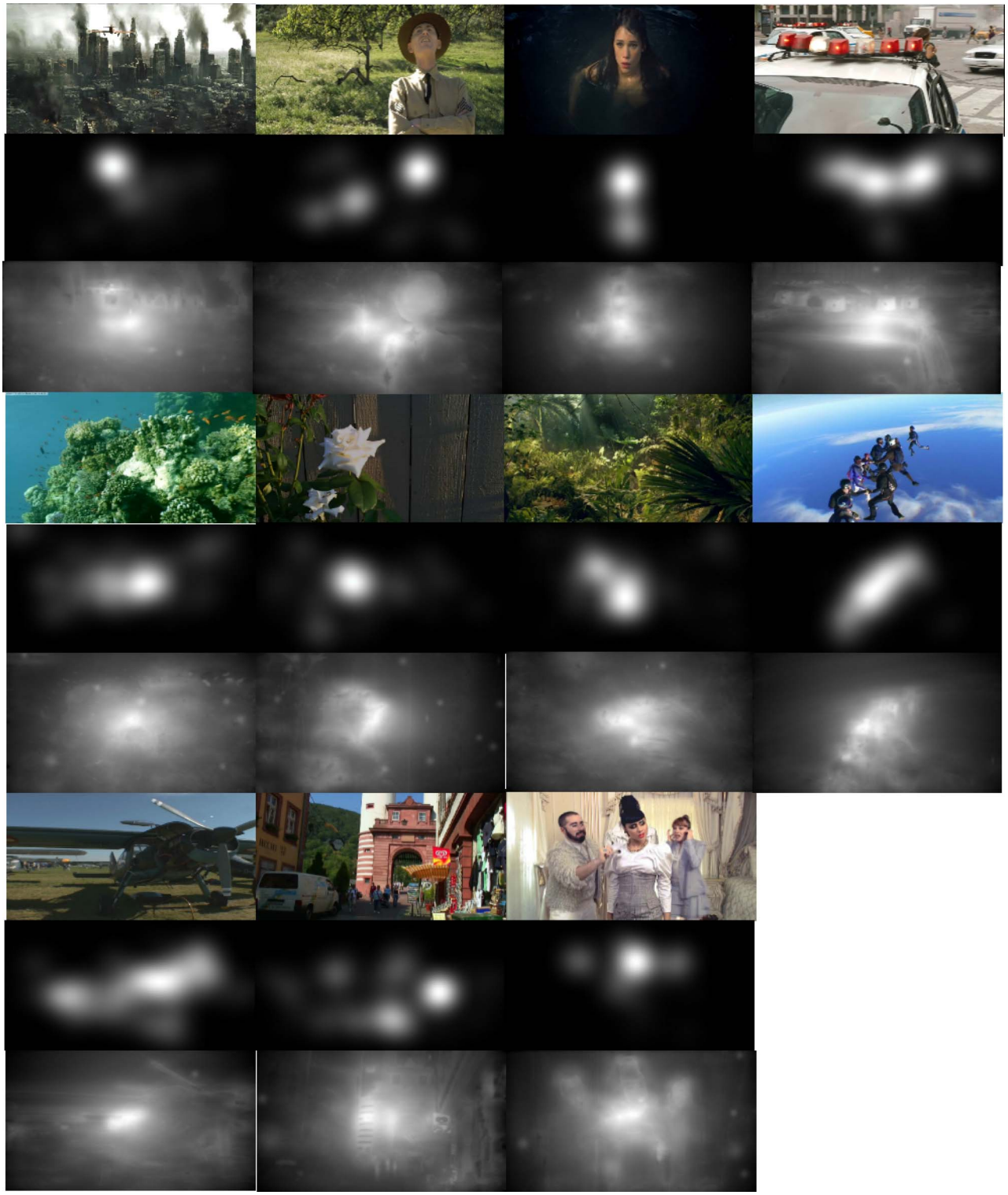

Figure 8. Original images (first row), human fixation density maps (second row), and our predicted saliency maps (third row) in the NCTU-3DFixation database.

Performance on the 3DGaze database

Wang et al. (2013) created and published an eye-tracking database containing 18 stereoscopic images with their associated disparity maps and the eye-movement data for both eyes. We applied our saliency model to these 18 images; the resultant AUC and PLCC metrics are shown in Table 3, and some predicted saliency maps are shown in Figure 11. Although our proposed model has better performance,

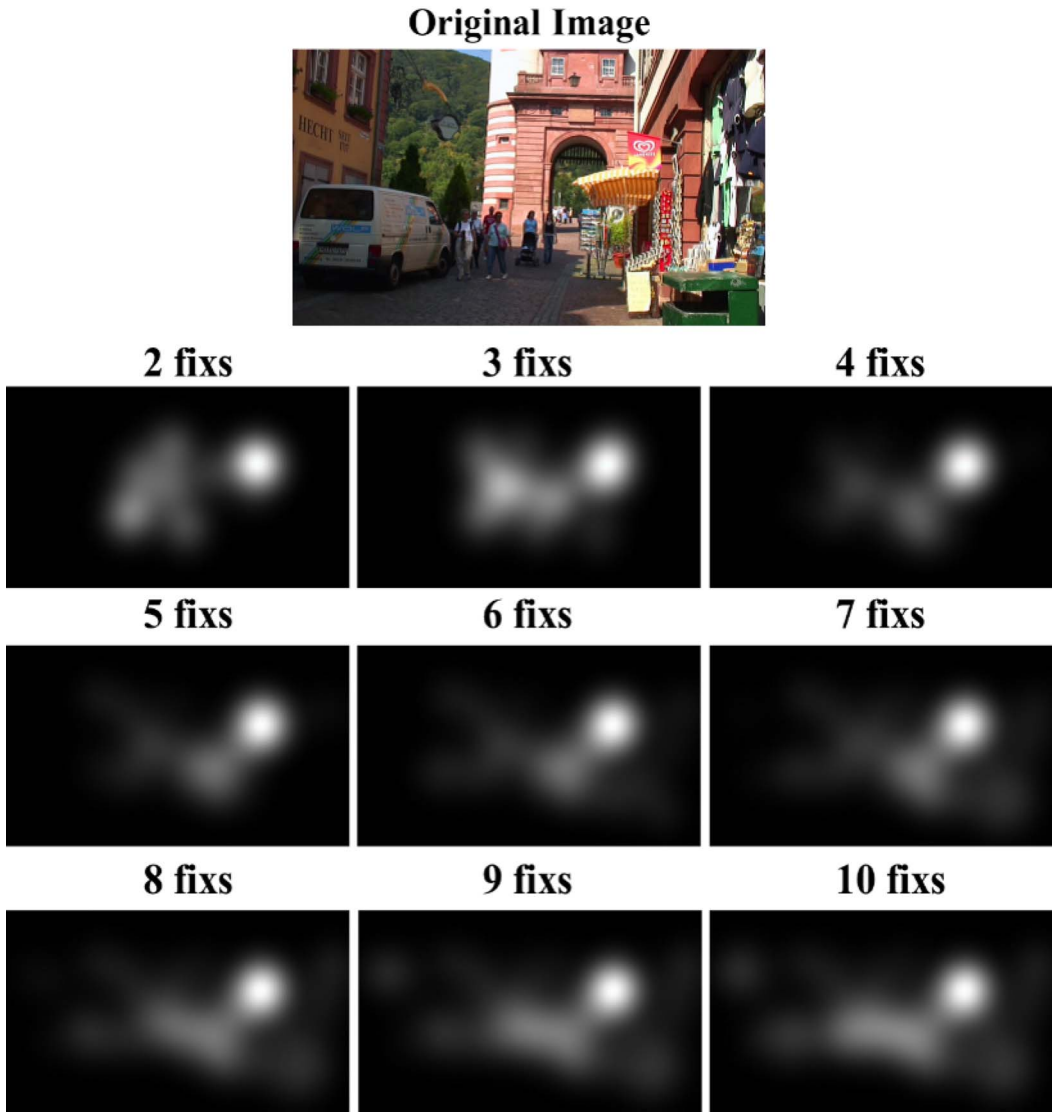

Figure 9. The human fixation density map varies with the number of fixations.

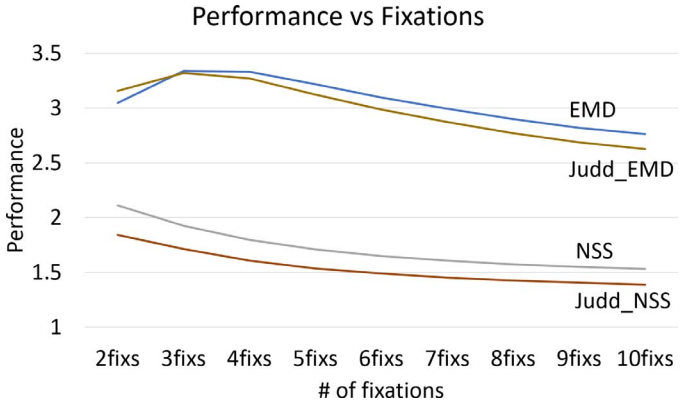

Figure 10. NSS and EMD metrics in our model and Judd et al.'s (2009) model based on the number of fixations.

it is not as good as that on the NCTU-3DFixation database. This is because some of their test images contain text or symbols, which are hard to predict by using simple feature extractors, but humans would focus on reading the text in a scene. For example, there are some letters in the second image of Figure 11, marked by red rectangles. Almost all the human observers focused on these letters. Another example is the third image; viewers tended to pay attention to the two letters on the stone in the lower right corner.

Also, as discussed earlier, the fixation ground truth map was created based on the observed eye-tracking data; however, different procedures for creating the ground truth map may lead to different ground truth maps. Although the same database is used in comparison, the procedure used by Wang et al. (2013) for creating the ground truth map is somewhat different from ours and thus, in certain cases, the ground truth

Model		AUC	PLCC
Proposed model		0.742	0.542
Wang et al.'s (2013) model	Itti's model (1998)	0.656	0.356
	Bruce's model (2009)	0.675	0.424
	Hou's model (2007)	0.670	0.410

Table 3. Performance comparison on Wang's dataset (Wang et al., 2013).

maps are different. Figure 12 depicts some extreme cases in which two ground truths (fixation density maps) show significant differences. For the first scene (Figure 12, left), our ground truth map clearly focuses on the stone in the lower right corner, which has two letters on it. In contrast, the ground truth map provided by Wang et al. focuses on the center of the test image. The same phenomenon can be found in the second scene (Figure 12, right). This difference may be due to the displacement compensation added by Wang et al. in creating the gaze point map. In this approach, a displacement, which can be horizontal or vertical, is

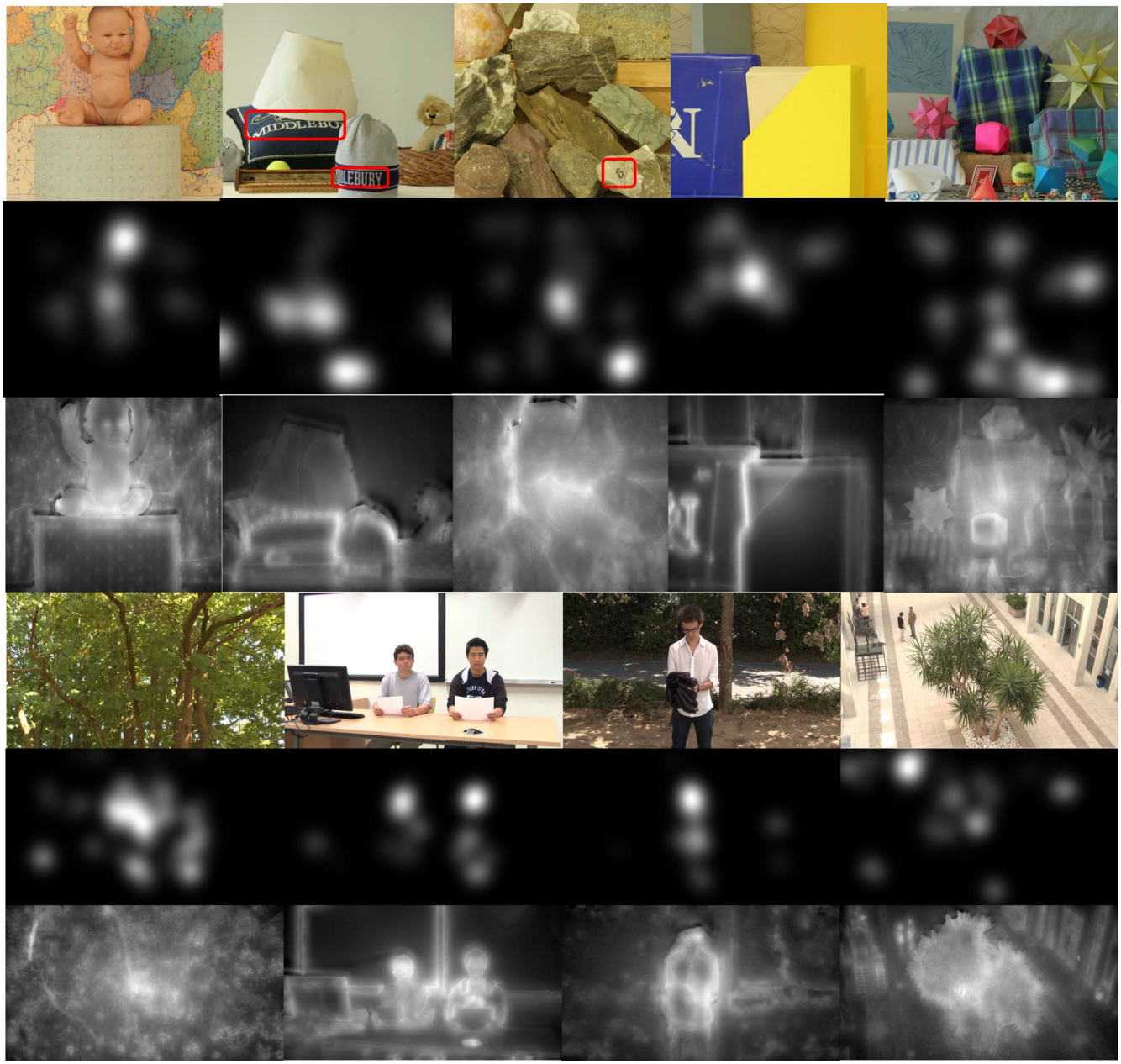

Figure 11. Examples of original images (first row), fixation density maps (second row), and our predicted saliency maps (third row) in the 3DGaze database.

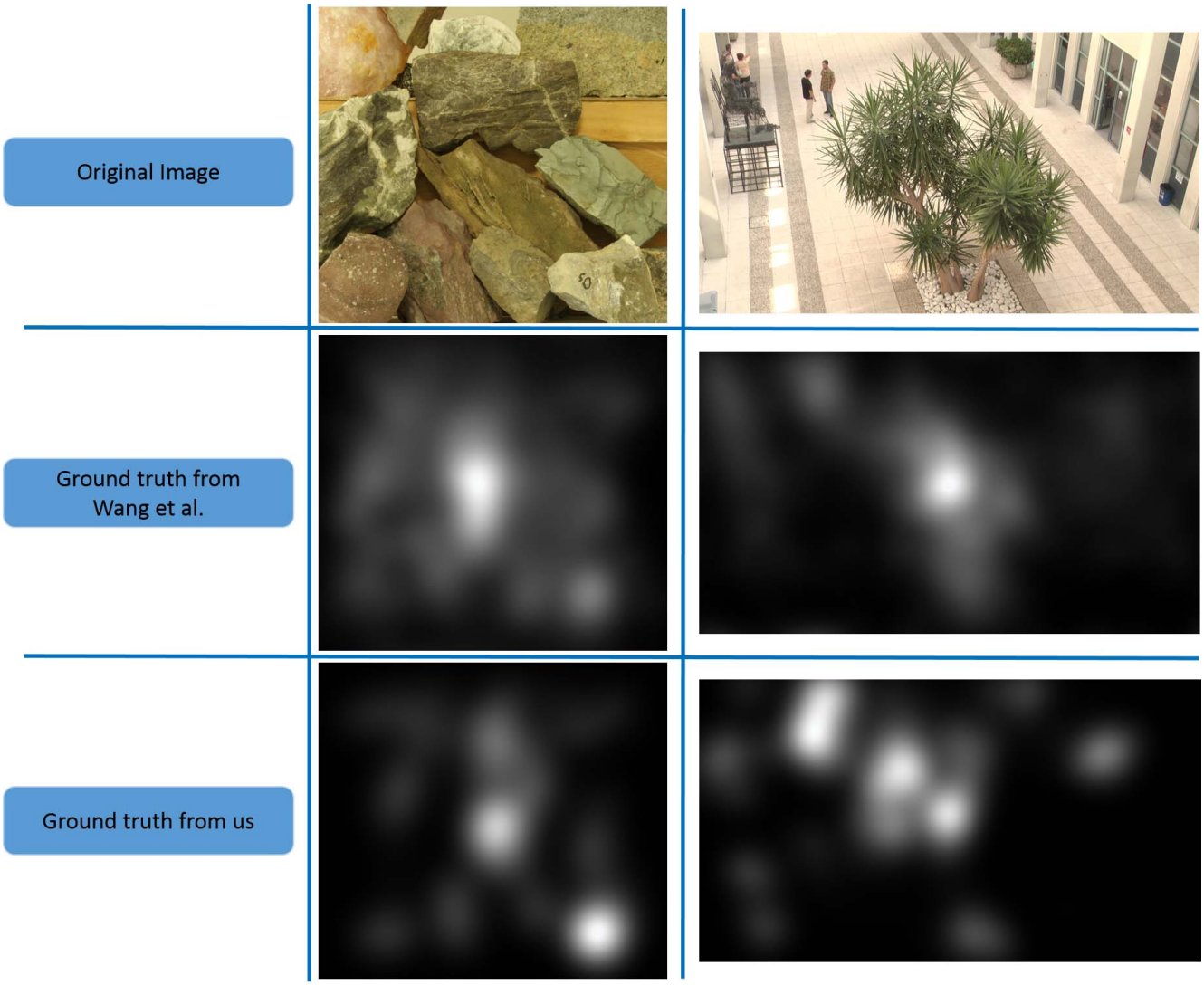

Figure 12. Original image and fixation density maps from Wang et al. (2013) and our fixation density maps.

added to the coordinates of each right-eye gaze point. The displacement of each gaze point comes from the right-to-left disparity map. In contrast, we believe that the disparity is mainly along the horizontal direction. Therefore, only horizontal displacement is used in our procedure, and the results seem to be generally closer to the expectations.

Discussion

Impact of each feature

We collected a number of features in the literature to predict the saliency map. At the end of the training process, our model was a linear combination of these feature values (maps). Therefore, the weight associated with a specific feature indicates the importance of that feature in predicting fixation. Figure 13 shows the weight distribution of all features used. For example, if a pixel has high values in both the center and the face feature maps, it means that there is a face at the center of the image and, thus, a human would likely pay attention to it. Some of the features we used (e.g., color, orientation, face, people, and center bias) were extremely important in our saliency prediction model. Which features were more essential than the others? It is possible that the weight of feature A is low because it coexists with features B and C. Thus, feature A may become one of the more essential features if features B and C are dropped. It is certainly a useful exercise to identify a minimum set of features that can provide good saliency map prediction. One advantage of the feature space dimension reduction is that it reduces computational time and memory.

For the purpose of finding the essential feature set, we selected several indispensable features individually

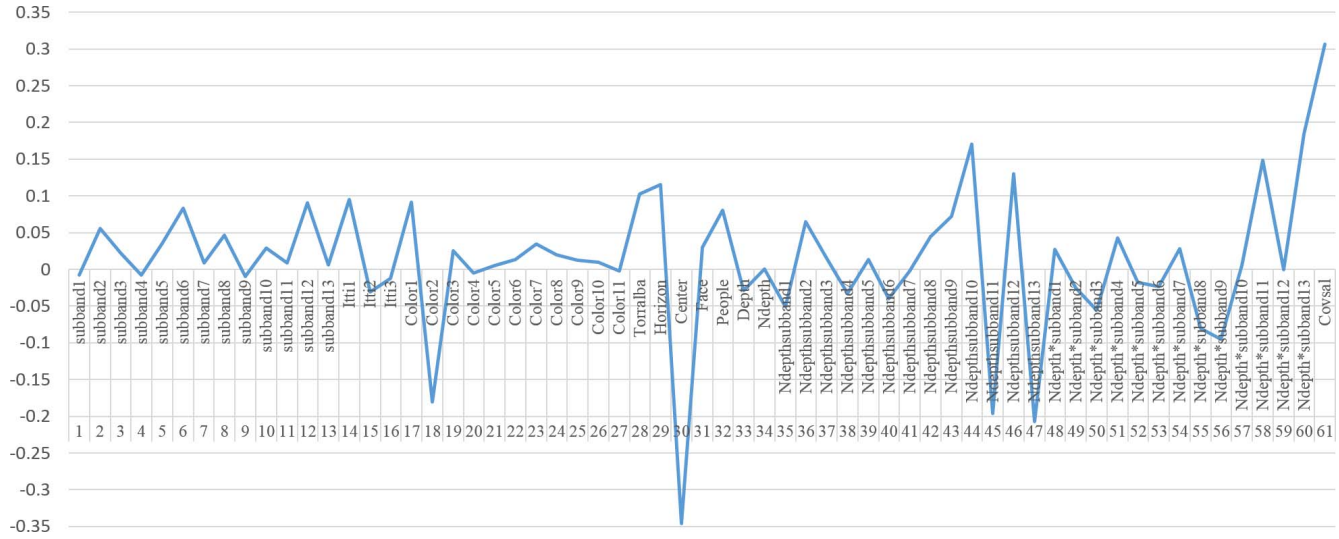

Figure 13. The weight distribution of all the low-, mid-, and high-level features. Refer to Figure 7 for the corresponding feature maps.

and tested their saliency map prediction performance. Then, we added features or replaced a few others and tested again. We could not exhaustively try all possible combinations of our features. The selected combinations of feature sets and their performance are summarized in Tables 4 and 5.

Note that we included the shuffled AUC (sAUC; first introduced by L. Zhang et al., 2008) to test the impact of various features. The main advantage of using the sAUC is that it emphasizes the off-center information and favors the true positives. The sAUC value of the center feature is near 0.5, which is consistent with the MIT saliency benchmarks (Judd et al., 2012).

Each column in Tables 4 and 5 is the set of selected features used together to predict the saliency map. The center bias feature alone provides a PLCC value of 0.542, an AUC value of 0.780, and an sAUC value of 0.507. This result is consistent with many reports that the center bias is the most essential feature in saliency map estimation. Thus, it is always included in our test feature sets. Table 4 examines each feature separately (together with the center bias). Statistically, the orientation and the color features are very influential on the performance. Another very important feature is CovSal. If we pick up only CovSal and the center bias, their performance is quite close to the best result we are aware of, such as Judd et al. (2009). In a way, the contribution of the orientation and the color feature set overlaps a lot with that of the CovSal feature. That is, if one set is selected, the other does not add much further improvement. Given the center bias, the rest of the features do not seem to increase the performance drastically. Note that the CovSal features contribute little to the sAUC score. This may be due to the final processing step in the CovSal model, which combines all the saliency maps at different scales by using the

Orientation		*		*					
Color			*	*					
Itti					*				
Horizon						*			
Center	*	*	*	*	*	*	*	*	*
Face							*		
People							*		
Depth								*	
CovSal									*
AUC	0.779	0.815	0.804	0.823	0.799	0.788	0.787	0.795	0.810
sAUC	0.507	0.595	0.562	0.594	0.544	0.511	0.522	0.538	0.529
PLCC	0.538	0.595	0.571	0.604	0.562	0.569	0.551	0.564	0.657
NSS	1.022	1.285	1.180	1.319	1.191	1.090	1.082	1.110	1.455
Similarity	0.529	0.539	0.529	0.537	0.533	0.540	0.531	0.515	0.625
EMD	2.981	2.628	2.755	2.596	2.816	2.992	2.895	2.883	2.727

Table 4. Impact of each feature together with center bias.

Orientation		*			
Color		*			
Itti			*		
Horizon	*	*	*	*	*
Center	*	*	*	*	*
Face	*	*	*	*	*
People	*	*	*	*	*
Depth				*	
CovSal					*
AUC	0.795	0.827	0.808	0.808	0.813
sAUC	0.533	0.610	0.563	0.556	0.538
PLCC	0.584	0.627	0.588	0.597	0.671
NSS	1.141	1.352	1.253	1.202	1.508
Similarity	0.545	0.546	0.543	0.525	0.632
EMD	2.930	2.607	2.805	2.793	2.678

Table 5. Impact of high-level features.

multiplication operation. This likely amplifies the importance of the high salient regions and degrades the rest, which lowers the sAUC score.

In addition, we examined the performance with only the orientation, color, depth, and center features. We selected 12 orientation, color, and depth features with high weighting values (numbers 6, 12, 17, 18, 30, 44, 45, 46, 47, 56, 58, and 60 in Figure 13) and reran the training procedure; the results are shown in Table 6. Note that the dimensionality has been significantly reduced, but its performance is still quite close to the case with all 61 features.

As discussed earlier, the steerable pyramid filter features on the depth map are introduced to improve the depth predictability. Figure 13 shows a specific weight pattern for features numbered from 44 to 47, which are associated with the steerable pyramid filters. These four weightings of different orientation decompositions indicate how the steerable pyramid filters extract the structural information from the depth map and how they are used to identify the salient objects. If a human observer is given only the depth image (instead of color image), he or she can still point out the strong salient objects in a scene, such as a human head, based on their depth map shapes. Thus, a good model should also be able to make proper use of the depth information. The weights in Figure 13 are trained by SVM on our collected data set.

Model	AUC	PLCC	NSS	Similarity	EMD
Proposed model with only highest weighting features	0.824	0.614	1.327	0.538	2.566

Table 6. Performance of the 12 highest weighting features.

Difficult cases and high-level features

We observed that some images are especially easy to predict, as shown in Figure 14. These images generally have a simple-texture background and contain only one or two interesting objects. On the other hand, our saliency model performs poorly on some other images. After careful examination, we observed that in most of these images most interesting objects are located at the border of the image, and thus the center bias fails. Two examples are shown in Figure 15. Two people and the monkey statue in the first image are located in the left and right corners. Even though our model can successfully recognize the two faces in the left bottom corner, the center bias feature still dominates the saliency map. The same phenomenon appears on the second sample image, in which a group of people is located in the lower left corner.

Indeed, many studies have revealed that faces and people in a picture attract human attention. However, in our experiments, these two features did not show much improvement over the center bias alone. This owes to two reasons. One is that the percentage of test pictures featuring faces and people is rather low. The other is that the detector used to identify their locations is not very reliable. Misidentification happens from time to time. One example is shown in Figure 7. There are three faces and three people in the original image; however, the face detector misidentified some regions as faces, and the people detector failed on the people in the center and the left side of the image. Examining the predicted saliency map, we concluded that the face and people features are critical to the specific cases in which these subjects show up but the other features do not assign large values to them. Because these cases are few, the average result is not very significant. In conclusion, to enhance the accuracy of saliency map estimation on every image, reliable face and people features are

Figure 14. Easy-to-predict images. First row: original images. Second row: fixation density maps. Third row: predicted saliency maps.

Figure 15. Difficult cases in which our saliency model works poorly. First row: original images. Second row: fixation density maps. Third row: predicted saliency maps.

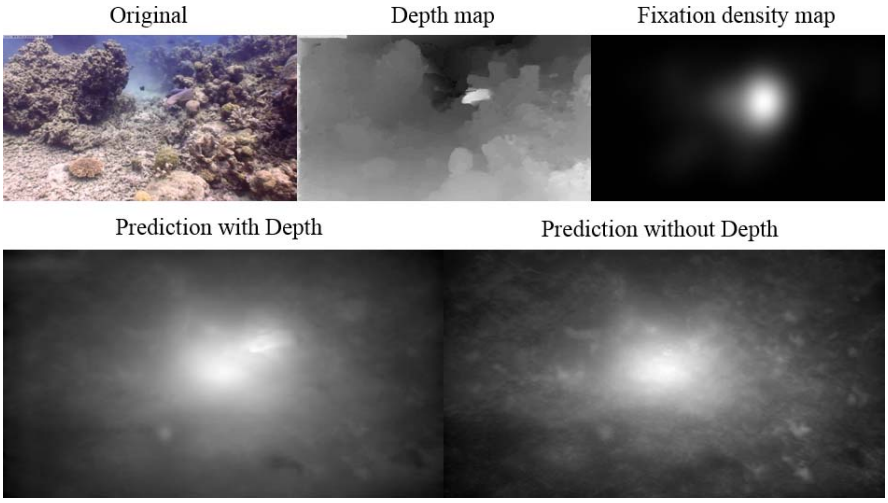

Figure 16. A case in which depth information is critical: complicated texture.

needed. They are complementary to the other features in certain cases.

Depth feature

The depth feature alone in our 3D scenes provided an average performance improvement (in addition to the center bias). To further study the influence of depth features, we trained another saliency model without using any depth information. After comparing the results from these two models, we noticed that the depth features are essential for certain images. Two such examples are shown in Figures 16 and 17.

Figure 16 shows a fish swimming near the ocean floor. The complex background makes it hard for the model without depth information to locate the object (i.e., fish) that humans pay attention to. In Figure 17, the interesting objects (i.e., two people) are not located at the center. These two people stand in the left corner and are close to the camera. The depth feature helps to allocate them and, consequently, the accuracy of the saliency model is significantly improved in this case. Thus, the depth feature is complementary to the other features in certain cases and can improve the saliency prediction quite a lot in these cases.

In Table 5, we include the face and people features in the basic feature set, which comprises the horizon feature and the center bias. We included additional features to test their performance. This comparison attempted to find the impact of each low-level feature. Similar to the observations from Table 4, the orienta-

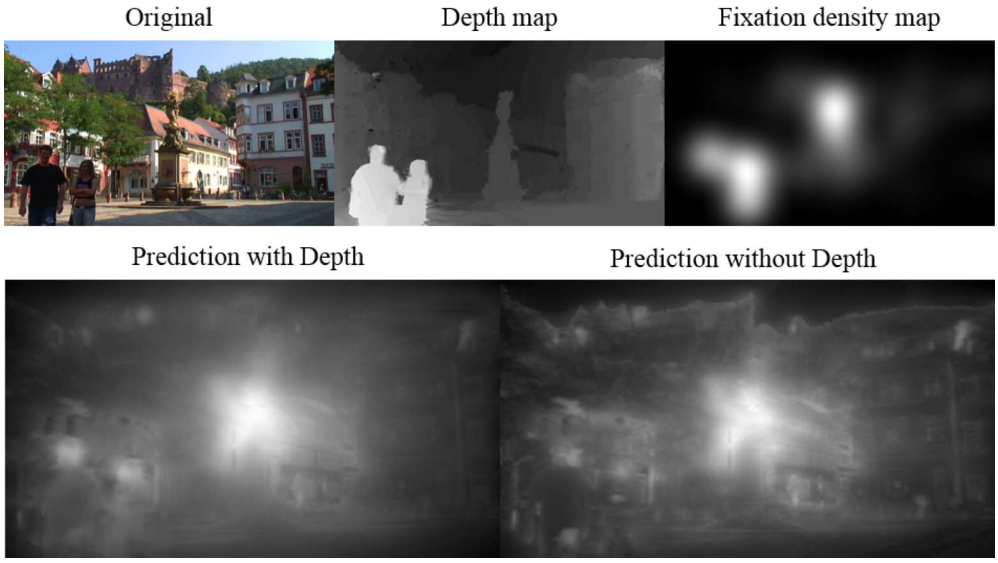

Figure 17. Another case in which depth information is critical: The interesting objects are away from the center.

tion and color features are very useful, and statistically the CovSal feature provides nearly the most improvement. However, the computation time of orientation features is only about 1 s and that of color features is about 8 s, while the computation time of the CovSal features is about 20 s per image. To save computation, a set of well-tuned weightings of the orientation and color features can achieve a comparable result.

Conclusions

This paper first describes the NCTU-3DFixation data set, which includes 475 3D images along with their depth maps and the eye-fixation data. We believe that this data set should be beneficial to the 3D visual attention research community, especially because its size is sufficient for training a learning-based saliency model. By applying the SVM algorithm to our data set, the best weights for low-, mid-, and high-level features were derived. The learning-based model was tested on two data sets, and the performance indicated that our model outperforms the known 2D and 3D models. After analyzing the data sets and comparing the feature weights in our model, we observed that the difference between watching 2D and 3D content is generally not very significant. However, there are some differences in the first a few fixations. Similar to the face and people features, the depth features are critical for certain difficult images. We also examined the performance of each feature and concluded that together with the basic center bias feature, the orientation and color features with their well-trained weights can achieve nearly the best performance. To facilitate the use of our 3D fixation data set, we released the data set together with the MATLAB codes for raw data extraction, visualization, SVM training, and evaluation methods on our website.

Keywords: visual attention, saliency map, depth saliency, eye-fixation database

Acknowledgments

The authors thank Chen-Chao Tao for offering his equipment and professional advice. Our research would not have been possible without his generous help and support. This work was supported in part by the NSC, Taiwan, under Grants NSC 101-2221-E-009-136 and NSC 102-2218-E-009-003 and by the Aim for the Top University Project of National Chiao Tung University, Taiwan.

Commercial relationships: none.

Corresponding author: Hsueh-Ming Hang.
Email: hmhang@mail.nctu.edu.tw.
Address: Department of Electronics Engineering, National Chiao-Tung University, Hsinchu, Taiwan.

References

Bruce, N. D., & Tsotsos, J. K. (2005). An attentional framework for stereo vision. In *2nd Canadian conference on computer and robot vision* (pp. 88–95). Piscataway, NJ: IEEE. doi:10.1109/CRV/2005.13.

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*(3):5, 1–24, http://www.journalofvision.org/content/9/3/5, doi: 10.1167/9.3.5. [PubMed] [Article]

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9*(12):10, 1–15, http://www.journalofvision.org/content/9/12/10, doi:10.1167/9.12.10. [PubMed] [Article]

Chamaret, C., Godeffroy, S., Lopez, P., & Le Meur, O. (2010). Adaptive 3D rendering based on region-of-interest. In A. Woods et al. (Eds.), *Proceedings of SPIE 7524, stereoscopic displays and applications*, XXI, 75240V. San Francisco, CA: SPIE. doi:10.1117/12.837532.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 27.

Didyk, P., Ritschel, T., Eisemann, E., Myszkowski, K., & Seidel, H. P. (2011). A perceptual model for disparity. *ACM Transactions on Graphics, 30*(4), 96, doi:10.1145/2010324.1964991.

Engelke, U., Liu, H., Wang, J., Le Callet, P., Heynderickx, I., Zepernick, H., & Maeder, A. (2013). Comparative study of fixation density maps. *IEEE Transactions on Image Processing, 22,* 1121–1133, doi:10.1109/TIP.2012.2227767. [PubMed]

Erdem, E., & Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision, 13*(4):11, 1–20, doi:10.1167/13.4.11. [PubMed] [Article]

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *IEEE conference on computer vision and pattern recognition* (pp. 1–8). Piscataway, NJ: IEEE. doi:10.1109/CVPR.2008.4587597.

Häkkinen, J., Kawai, T., Takatalo, J., Mitsuya, R., & Nyman, G. (2010). What do people look at when they watch stereoscopic movies? In A. Woods et al. (Eds.), *Proceedings of SPIE 7524, stereoscopic displays and applications*, XXI, 75240E. San Francisco, CA: SPIE. doi:10.1117/12.838857.

Huynh-Thu, Q., & Schiatti, L. (2011). Examination of 3D visual attention in stereoscopic video content. In B. Rogowitz & T. Pappas, (Eds.), *Proceedings of SPIE 7865, human vision and electronic imaging*, XVI, 78650J. San Francisco, CA: SPIE. doi:10.1117/12.872382.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506, doi:10.1016/S0042-6989(99)00163-7. [PubMed]

Jansen, L., Onat, S., & König, P. (2009). Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision, 9*(1):29, 1–19, http://www.journalofvision.org/content/9/1/29, doi:10.1167/9.1.29. [PubMed] [Article]

Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. *MIT-CSAIL-TR-2012-001.*

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *IEEE 12th international conference on computer vision* (pp. 2106–2113). Piscataway, NJ: IEEE.

Kim, H., Sanghoon, L., & Bovik, A. C. (2014). Saliency prediction on stereoscopic videos. *IEEE Transactions on Image Processing, 23,* 1476–1490, doi:10.1109/TIP.2014.2303640. [PubMed]

Lang, C., Nguyen, T. V., Katti, H., Yadati, K., Kankanhalli, M., & Yan, S. (2012). Depth matters: Influence of depth cues on visual saliency. In *Computer vision–ECCV* (pp. 101–115). Berlin Heidelberg: Springer. doi:10.1107/978-3-642-33709-3_8.

Li, Z., & Itti, L. (2008). Visual attention guided video compression. *Journal of Vision, 8*(6):772, http://www.journalofvision.org/content/8/6/772, doi:10.1167/8.6.772. [Abstract].

Liu, Y., Cormack, L. K., & Bovik, A. C. (2010). Dichotomy between luminance and disparity features at binocular fixations. *Journal of Vision, 10*(12):23, 1–17, http://www.journalofvision.org/content/10/12/23, doi:10.1167/10.12.23. [PubMed] [Article]

Maki, A., Nordlund, P., & Eklundh, J. O. (1996). A computational model of depth-based attention. *Proceedings of the 13th International Conference on Pattern Recognition, 4,* 734–739, doi:10.1109/ICPR.1996.547661.

Oliva, A., & Torralba, A. (2001). The steerable pyramid: Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145–175, doi:10.1023/A:1011139631724.

Ouerhani, N., & Hugli, H. (2000). Computing visual attention from scene depth. *IEEE 15th International Conference on Pattern Recognition, 1,* 375–378, doi:10.1109/ICPR.2000.905356.

Pele, O., & Werman, M. (2009). Fast and robust Earth mover's distances. In *IEEE 12th international conference on computer vision* (pp. 460–467). Piscataway, NJ: IEEE.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45,* 2397–2416, doi:10.1016/j.visres.2005.03.019. [PubMed]

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*(1), 3–25. [PubMed]

Potapova, E., Zillich, M., & Vincze, M. (2011). Learning what matters: Combining probabilistic models of 2D and 3D saliency cues. *Computer Vision Systems, 6962,* 132–142, doi:10.1007/978-3-642-23968-7_14.

Psychology Software Tools, Inc. [E-Prime 2.0]. (2012). Retrieved from http://www.pstnet.com.

Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., & Chua, T. S. (2010). An eye fixation database for saliency detection in images. *Proceedings of the 10th European Conference on Computer Vision, 6314,* 30–43, doi:10.1007/978-3-642-15561-1_3.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research, 39,* 3157–3163, doi:10.1016/S0042-6989(99)00077-2. [PubMed]

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The Earth movers distance as a metric for image retrieval. *International Journal of Computer Vision, 40*(2), 99–121, doi:10.1023/A:1026543900054.

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. *International Conference on Image Processing, 3,* 444–447, doi:10.1109/ICIP.1995.537667.

Stankiewicz, B. J., Anderson, N. J., & Moore, R. J. (2011). Using performance efficiency for testing and optimization of visual attention models. In S. Farnand & F. Gaykema, (Eds.), *Proceedings of SPIE 7867, image quality and system performance*, VIII, 78670Y. San Francisco, CA: SPIE. doi:10.1117/12.872388.

Tanimoto, M. (2012). FTV: Free-viewpoint television. *Signal Processing: Image Communication, 27,* 555–570, doi:10.1016/j.image.2012.02.016.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14):4, 1–17, http://www.journalofvision.org/content/7/14/4, doi:10.1167/7.14.4. [PubMed] [Article]

Viola, P., & Jones, M. (2001). Robust real-time object detection. *International Journal of Computer Vision, 4,* 34–47.

Wang, J., Da Silva, M. P., Le Callet, P., & Ricordel, V. (2013). Computational model of stereoscopic 3D visual saliency. *IEEE Transactions on Image Processing, 22*(6), 2151–2165, doi:10.1109/TIP.2013.2246176. [PubMed]

Wang, J., Le Callet, P., Tourancheau, S., Ricordel, V., & Da Silva, M. P. (2012). Study of depth bias of observers in free viewing of still stereoscopic synthetic stimuli. *Journal of Eye Movement Research, 5*(5), 1–11, doi:10.1109/CRV.2005.13.

Yarbus, A. L. (1967). *Eye movements and vision*. New York, NY: Plenum Press.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7):32, 1–20, http://www.journalofvision.org/content/8/7/32, doi:10.1167/8.7.32. [PubMed] [Article]

Zhang, Y., Jiang, G., Yu, M., & Chen, K. (2010). Stereoscopic visual attention model for 3D video. *Advances in Multimedia Modeling, 5916,* 314–324, doi:10.1007/978-3-642-11301-7_33.

Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision, 11*(3):9, 1–15, http://www.journalofvision.org/content/11/3/9, doi:10.1167/11.3.9. [PubMed] [Article]