Chapter 8
Sensing and Controlling Human Gaze in Daily Living Space for Human-Harmonized Information Environments

Yoichi Sato, Yusuke Sugano, Akihiro Sugimoto, Yoshinori Kuno and Hideki Koike

Abstract This chapter introduces new techniques we developed for sensing and guiding human gaze non-invasively in daily living space. Such technologies are the key to realize human-harmonized information systems which can provide us various kinds of supports effectively without distracting our activities. Toward the goal of realizing non-invasive gaze sensing, we developed gaze estimation techniques, which requires very limited or no calibration effort by exploiting various cues such as spontaneous attraction of our visual attention to visual stimuli. For shifting our gaze to desired locations in a non-disturbing and natural way, we exploited two approaches for gaze control: subtle modulation of visual stimuli based on visual saliency models, and non-verbal gestures in human-robot interactions.

Keywords Appearance-based gaze sensing · Calibration-free gaze estimation · Visual saliency · Gaze guidance

Y. Sato (✉)
Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo, Japan
e-mail: ysato@iis.u-tokyo.ac.jp

Y. Sugano
Max Planck Institute for Informatics, 66123 Saarbrücken, Germany
e-mail: sugano@mpi-inf.mpg.de

A. Sugimoto
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
e-mail: sugimoto@nii.ac.jp

Y. Kuno
Saitama University, 255 Shimo-Okubo, Sakura-ku, Saitama, Japan
e-mail: kuno@cv.ics.saitama-u.ac.jp

H. Koike
Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo, Japan
e-mail: koike@cs.titech.ac.jp

T. Nishida (ed.), *Human-Harmonized Information Technology, Volume 1*,
DOI 10.1007/978-4-431-55867-5_8

8.1 Introduction

When we interact with other people to support or collaborate with them in various situations in our daily lives, we always pay close attention to what other people are attended to. With such awareness of other people's attention, we are able to decide when to interrupt and, if necessary, provide support to other people without distracting them. The awareness of attention is also expected to play a key role in realizing natural human-machine interactions. Therefore, a human-harmonized information system should be capable of sensing our attention in a non-invasive manner, so that the system can provide us various supports without distracting our activities. Moreover, we argue that, not only sensing human attention, human-harmonized information systems should be able to guide our attention naturally at appropriate timing to a desired location.

This motivated us to develop novel techniques for sensing and guiding human attention in a non-invasive manner. This chapter describes the outcome of our CREST project toward the goal, more specifically, remote gaze sensing methods that require less or no calibration effort (Sect. 8.2), gaze guidance based on visual saliency and its application (Sect. 8.3), and gaze guidance by a humanoid robot with non-verbal behaviors (Sect. 8.4).

8.2 Human Gaze Sensing in Daily-Life Environments

The goal of gaze estimation is to determine where a person is looking at. While we need to rely on head orientation as a rough indicator of human gaze in far-distance settings such as surveillance camera systems [5], observing eyes is the most direct way to infer the target person's gaze. Because of its wide variety of application fields ranging from scientific studies to practical applications, various eye gaze estimation techniques have been studied for many years [16]. However, existing techniques still have some critical limitations affecting the easiness of use, which is one of the most important properties required for a gaze estimation technique in daily-life environments.

In this section, we introduce our research attempts on appearance-based gaze estimation. In Sect. 8.2.1, we discuss the core technique of appearance-based method that require only a remote camera to estimate eye gaze directions. In Sect. 8.2.2, we further introduce approaches to make appearance-based methods *calibration-free*, working without any personal calibration actions.

8.2.1 Appearance-Based Gaze Estimation

In general, remote gaze estimation techniques have a great advantage that they do not require users to wear special devices. Existing remote gaze techniques can be

mainly categorized into either model-based or appearance-based approaches. Model-based approaches estimate pose of a geometric eyeball model using high-resolution eye cameras and additional light sources. While they can achieve relatively higher accuracy and are taken in commercial eye trackers, it often requires specialized hardware such as high-resolution close-up cameras and additional light sources. In contrast, appearance-based approaches only require eye images and thus have an advantage when only low-resolution images are available as input.

The biggest limitation of appearance-based gaze estimation is that it requires relatively larger amount of calibration data to establish the estimation function. In this section, we introduce the adaptive linear regression (ALR) method to reduce the required number of training samples [30]. This can be realized by introducing more effective regression algorithms via ℓ^1-minimization that can work with sparse training samples.

8.2.1.1 Adaptive Linear Regression

Eye Feature Extraction

Existing appearance-based methods generate the eye image feature from a captured image by raster scanning all its pixels, thus the typical feature dimensionality reaches several thousand [46] or even higher (e.g. edge map is added in [51]). On one hand, a high-dimensional feature keeps all the information in the image. On the other hand, since gaze directions have only two degrees of freedom, such high-dimensional features are highly redundant. Moreover, actually captured eye regions can be of variant resolutions, and therefore, pixel-wise feature extraction faces the problem of inconsistent output dimensions.

We use a low-dimensional feature extraction method consisting of two steps: eye region alignment and feature generation. In the first step, the eye regions are accurately aligned for different eye images. The inner and outer eye corners from an anchor eye image are first detected by an edge filter. Then the eye corner regions are stored as image templates to match the eye corners of other eye images, and finally align the eye region.

In the feature generation step, once the aligned eye image region is obtained, it is further divided into 15 even subregions, as shown in Fig. 8.1. Then the feature vector

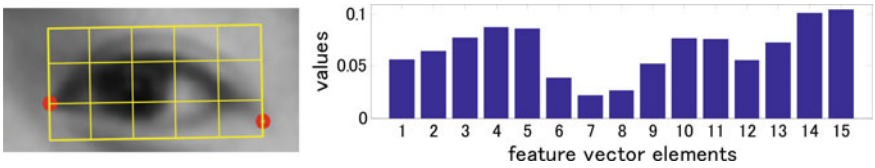

Fig. 8.1 Eye appearance feature extraction. *Left* Illustration of eye corner detection and 3×5 subregion division. *Right* Generated 15-D feature vector (Reproduced with permission of © 2014 IEEE [30].)

is generated by summarizing pixel values in each subregions and combining them to form a unique feature vector.

Eye Appearance Manifold

All the eye appearance features, which are extracted from the accurately aligned image regions, constitute a manifold in the high dimensional space. Since the eyeball movement has only two degrees of freedom, the manifold has an intrinsic dimensionality of close to two. To test this statement, we project all the features onto a 3-D space by PCA for visualization, as shown in Fig. 8.2. Several observations can be made here. First, the eye feature manifold can be approximated as a 2-D surface with most of its information accumulating inside the first two major dimensions. Second, the proposed 15-D feature well keeps more information in the first three major dimensions (Fig. 8.2c) than the pixel-wise extracted feature (Fig. 8.2b). Therefore, although the manifold in Fig. 8.2b seems smoother, the information can be hidden in other dimensions. Finally but most importantly, the projected features on the manifolds in Fig. 8.2b, c show a similar pattern to the gaze positions in Fig. 8.2a.

Local Regression via Interpolation

Following these observations, without any prior knowledge, unknown gaze positions can be found by using linear interpolation by assuming locality, i.e. limiting the interpolation within a sufficient small region centered by the unknown sample in the manifold. The existing methods [46] guarantee this locality assumption by obtaining dense training samples, from which they select some training samples with the smallest Euclidean distances from the unknown for interpolation. However, if the training samples are only sparsely collected, as shown in Fig. 8.2, the local linearity assumption cannot be satisfied. Therefore, problem with sparse training samples motivates our methods. Our idea is to adaptively find an optimal set of training samples that best reconstruct the test image linearly, and we show that by using the same

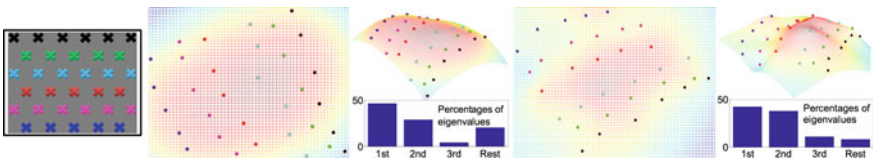

Fig. 8.2 2-D gaze space and eye appearance feature manifold. *Left* Illustration of 33 gaze positions on 2-D screen. *Middle* Projection of corresponding eye appearance manifold on 3-D space. The magnitudes of the eigenvalues are shown as percentages. *Right* Illustration of manifold projection for our proposed 15-D low dimensional feature. Notice the similarity between the gaze positions on the screen and the feature coordinates on the manifold (Reproduced with permission of © 2014 IEEE [30].)

linear combination, gaze estimation can be done accurately without requiring dense training data.

The idea is to take into consideration all the training samples at first, and then adaptively choose some of them in interpolation. The way to do the choosing is to use the ℓ^1-minimization. In particular, our interpolation can be written in a standard matrix multiplication form, where the weights tell how much each training sample contributes. If one weight is zero, it means the corresponding training sample is not selected, and vice versa. In addition, we put a constraint in our formulation so that the summation of all weights is always one, and we also have an ε to control the the interpolation error. These weights are optimized by using any of the efficient ℓ^1 solvers and then used to compute the unknown gaze position via interpolation.

ℓ^1-minimization has been used in previous vision-related researches, among which the works by Wright et al. [52], Wagner et al. [49] and Tan et al. [47] share similarities to ours. However, our work essentially differs in some aspects. First, they apply face recognition, which is a classification problem, while we handle a typical regression problem. In this sense, our focus is not discriminability but accurate estimation, and we newly show the effectiveness of ℓ^1-optimization-based method in handling a regression problem. Second, Wright et al. [52] also introduce an error term ε without mentioning how to determine its value. However, we find that the ε value is crucial to our solution and should be carefully chosen, while it is not the case for classification problems. As a result, we dynamically optimize ε by checking whether the ℓ^1 norm of our weights is equal to one. Another difference is that [52] assumes sparsity in reconstruction errors while we assume fewer supporting training samples for optimal representation for a query image.

Evaluation

We evaluate the estimation accuracy of the basic ALR method in this section. Details of the experimental setups and approaches include:

- **Training samples**. The training samples were sparsely collected in four sets with totals of 9, 18, 23, and 33.
- **Test samples**. For each training set, nearly 100 test samples were collected whose gaze positions were randomly chosen on the screen.
- **Dataset size**. The four sets of training/test samples were collected for each of the seven subjects.
- **Fixed head pose**. A chin rest was used to help stabilize the users' heads.
- **Eye image alignment**. With fixed head poses, the eye regions were directly cropped from the same region for all images.
- **Feature**. In our case, 15-D features were extracted with 3×5 subregions as described above. While for other methods in comparison, different features were generated and used as they proposed.

We compare the proposed ALR method with some latest appearance-based methods in terms of gaze position estimation accuracy. These methods were proposed by

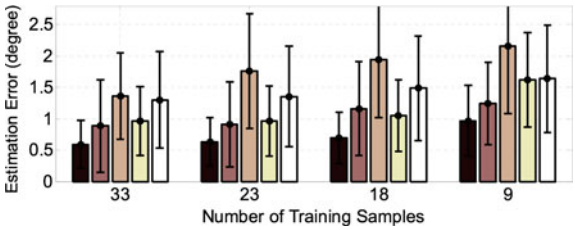

Fig. 8.3 Comparison of average gaze estimation results. For each number of training samples, results of ALR, local region [46], PCA+GPR [34], HOG+SVR [31], and CSLBP+GPR [28] are shown from *left* to *right* (Reproduced with permission of © 2014 IEEE [30].)

using different feature descriptors as and regression techniques [28, 31, 34, 46]. We use the same dataset to test their performances. As a result, their average estimation errors and their standard deviations are given in Fig. 8.3. The proposed method shows the highest estimation accuracy in different experimental conditions. In general, it achieved estimation accuracies of better than 1°. As for other methods, although they use more complex descriptors and different regression techniques, their accuracies are clearly not as good as ours in the conditions of sparse training samples.

8.2.1.2 Allowing Free Head Motion

Another major limitation of the appearance-based methods lies in that most of them assume a fixed head pose. This limitation is hard to remove because the head motion has 6 degrees of freedom which must be handled by more training data. Therefore directly solving the problem requires a prohibitively large number of training samples.

To effectively solve this problem while significantly reduce the training cost, we also proposed an approach extending the ALR method [29]. In this method, we initially estimate the gaze direction under the fixed head pose as introduced above, and then apply a series of rotations to the head coordinate system so that the gaze direction error caused by geometric head rotation can be compensated. We further investigate the relationship between the changes of head directions and the biases of gaze estimations caused by eye appearance distortions, and propose an additional calibration process using a 5 s video clip. This achieves the average estimation accuracy of 2.4° under free head motion without using complex devices such as infrared/stereo cameras/lights and pan-tilt units.

8.2.2 Calibration-Free Gaze Estimation

One of the biggest limitation of existing gaze estimation techniques is that they require person-specific calibration. As discussed in the previous section, the calibration is in

general done by showing some gaze targets to the target user and acquiring eye images with ground-truth gaze directions. Since the calibration process requires an active participation of the user and makes natural gaze estimation impossible, it can bring a strong constraint on the application scenarios. Calibration drift is another issue that is shared among existing techniques, and their performances significantly decrease when the condition becomes different from the initial calibration setting. In this subsection, we introduce two different calibration-free gaze estimation approaches to address this problem.

8.2.2.1 Saliency-Based Auto Calibration

The key idea of the first approach [44] is to acquire ground-truth calibration information from the target user's natural behavior. This idea can be realized by using computational models of *visual saliency*.

It is known that humans can rapidly look at regions with high visual saliency, i.e., a region containing unique and distinctive visual features compared to the surrounding regions. Computational visual saliency models have been studied to mimic and understand this mechanism of visual attention, and can be used to estimate visual saliency maps in a bottom-up manner from images and videos (see [2] for a recent survey). The visual saliency maps computed from a video that the user is seeing tell us which region can attract more attention, and hence can be used as a probabilistic calibration data to learn the mapping from the eye images to the gaze points.

Our method takes a set of eye images recorded in synchronization with a video clip as the input, and automatically determines the relationship between the eye images and gaze directions. While existing methods require additional calibration data to estimate gaze positions of these eye images, in our method the input data can also serve as the calibration data.

The proposed method is illustrated in Fig. 8.4. Once the saliency maps are extracted from the input video frames, we aggregate the saliency maps based on the similarity of the eye images to produce more accurate gaze probability maps.

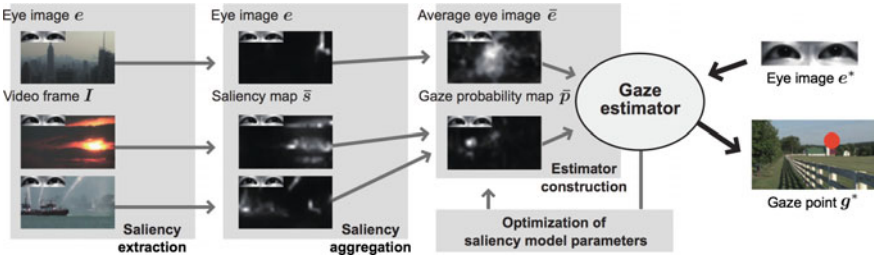

Fig. 8.4 Illustration of the saliency-based calibration approach. The method consists of mainly four steps: saliency extraction, saliency aggregation, estimator construction and saliency model optimization (Reproduced with permission of © 2013 IEEE [44])

The method then learns the relationship between the gaze probability maps and the eye images. In addition, a feedback scheme optimizes the feature weights used to compute the visual saliency maps. The feedback loop enables us to further strengthen the relationship between the gaze probability maps and the eye images.

Saliency Extraction

Our method adopts five low-level features and one high-level feature to compute the saliency maps. For low-level features, we use commonly-used feature channels, i.e., color, intensity, and orientations as the static features, and flicker and motion as dynamic features. Five saliency maps are computed from these features using the Graph-based Visual Saliency (GBVS) algorithm [17]. In addition to these low-level features, it is well known that humans tend to fixate on salient objects such as human faces. In order to capture this high-level saliency, we also compute a face channel-based saliency model [4] using a face detector. As a result, synchronized pairs of six saliency maps and eye images are produced. As can be seen in the examples shown in Fig. 8.4, saliency maps represent saliency values in the image space, and highly salient regions in the saliency map are likely to coincide with the actual gaze point.

Saliency Aggregation

Although the saliency maps can be correlated with gaze point distributions, their accuracy is insufficient for determining the exact gaze point locations on each map. We hence compute the probability distribution of the gaze point by aggregating the computed saliency maps. When we assume a fixed head position, there is a one-to-one correspondence between the ground-truth gaze points and the eye appearance. Therefore, by aggregating the saliency maps based on the similarity of the associated eye images, we can assume that the region around the actual gaze point has a sharp peak of saliency.

The eye images are first clustered according to their similarity scores, and then weighted means of the corresponding saliency maps are computed based on the similarity scores. The aggregated map can be used as the gaze probability map which represents the probability distribution of the gaze point more accurately.

Gaze Estimation

We then establish a mapping from the eye image to gaze points using the set of average eye images and corresponding gaze probability maps obtained in the previous step. Unlike standard calibration data, the gaze probability map only provides the probability distribution of the gaze point. Hence we approximate marginalization of the gaze probability with a Monte Carlo approximation, and apply a Gaussian process regression [37] to find the mapping function.

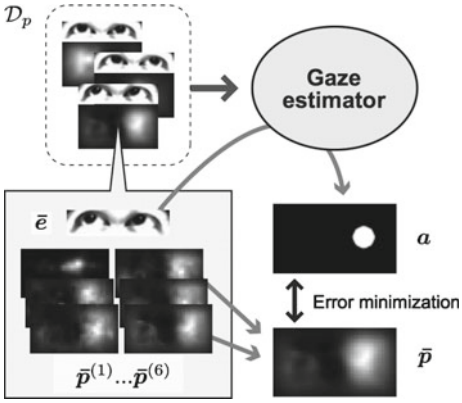

Fig. 8.5 Illustration of feature weight optimization. Feature weights are optimized by maximizing the similarity between estimated gaze positions and weighted sum maps (Reproduced with permission of © 2013 IEEE [44])

Feature Weight Optimization

In the first path, all six saliency features are independently aggregated, and the maps are linearly combined to produce the summed gaze probability map. In our method, the feature weights are further refined through an optimization loop as illustrated in Fig. 8.5. Once the gaze estimator is built, it can be used to estimate the gaze points from the average eye images. Using this data, our method optimizes the feature weights by minimizing the sum of the squared residuals between weighted sum maps and estimated gaze position maps. This maximizes the consistency between the gaze estimator and the saliency model.

Evaluation

In order to evaluate the performance of the proposed method, we used a set of 80 online video clips which include various types of video clips such as music videos and short films. We randomly extracted 30 s video sequences from each video source without an audio signal, and created four 10 min datasets A, B, C, D. Seven novice test persons are asked to watch all of them, and the estimation errors are evaluated using ground-truth positions obtained using a commercial gaze tracker. As a baseline, we also compare our method with a standard appearance-based gaze estimation method that uses an explicit calibration.

The estimation results are summarized in Table 8.1. Each row corresponds to the result using each dataset, where all 20 video clips are used for both the training and testing. The columns list the distance and angular errors of the proposed method and the calibrated appearance-based estimator. The overall average error was 39 mm ($\approx$3.5°), which is comparative to the calibrated estimator and sufficient for obtaining the regions of attention in images.

Table 8.1 Average error for each dataset

	Proposed method		Calibrated method	
Dataset	Error (mm)	Error (deg.)	Error (mm)	Error (deg.)
A	41 ± 26	3.7 ± 2.3	33 ± 15	3.0 ± 1.4
B	36 ± 23	3.3 ± 2.1	24 ± 13	2.2 ± 1.2
C	41 ± 25	3.7 ± 2.3	27 ± 15	2.5 ± 1.4
D	36 ± 25	3.3 ± 2.3	34 ± 16	3.1 ± 1.5
Average	$\mathbf{39 \pm 25}$	$\mathbf{3.5 \pm 2.3}$	30 ± 15	2.7 ± 1.4

The columns are the distance and angular estimation errors (average $\pm$ standard deviation) when using our method and the baseline method with an explicit calibration (Reproduced with permission of © 2013 IEEE [44])

8.2.2.2 Learning-Based Person-Independent Estimation

Another possible strategy for calibration-free gaze estimation is preliminary training a generic gaze estimator that can handle arbitrary users. If we have a large dataset that contains diverse people, head poses, and gaze directions, it is possible to train a person- and head pose-independent gaze estimation function. In this section, we introduce our method based on a multi-view gaze dataset [45].

In order to increase the head pose variation in the dataset, images are recorded by a multi-camera system and the view synthesis is performed via 3D reconstruction of eye regions. The gaze estimator is trained by an extension of random forests, and the best performance is achieved in person- and pose-independent, calibration-free gaze estimation from low-resolution images.

Dataset

For the purpose of learning a person- and pose-independent gaze estimation function, the training dataset must contain a dense samples in terms of persons, head poses, and gaze directions. In addition, accurate 3D positions of eyes and gaze targets in the camera coordinate system have to be provided as an annotation, and the coordinate system must be consistent across persons.

A total of 50 people participated in the data collection. As shown in Fig. 8.6 (left), eight cameras are attached to the frame of a monitor, and intrinsic and extrinsic camera parameters are calibrated beforehand. In order to obtain ground-truth gaze target positions displayed on the monitor, the 3D position of the monitor plane in the camera coordinate system is also calibrated. A chin rest was used to stabilize the head position located at 60 cm apart from the monitor, and participants were instructed to look at a visual target displayed on the monitor. The visual target were displayed at 160 regular grid positions in a random order. As a result, 160 (gaze directions) $\times$ 8 (cameras) $\times$ 50 (persons) images were recorded.

Fig. 8.6 Data collection. *Left* System configuration for data collection. *Right* Examples of reconstructed 3D eye region models (Reproduced with permission of © 2014 IEEE [45])

Data Annotation and 3D Reconstruction

The captured images are further annotated with facial landmarks. The locations of six facial landmarks are manually annotated on the first eight images for each person, and they are refined via a simple multi-view template matching on the rest of the images. These 3D landmark positions are further used to define head poses of the data.

We use a patch-based multi-view stereo algorithm [11] to reconstruct the 3D shapes from 8 multi-view images. The reconstructed 3D point cloud is further preprocessed by outlier removal and smoothing, and we then apply a Poisson reconstruction to obtain the 3D mesh of the eye region. The texture of the 3D mesh is computed using the mean of all source images. Figure 8.6 (right) shows examples of the reconstructed models.

Training Data Synthesis

Using the dataset, we take a learning-by-synthesis approach to person- and pose-independent gaze estimation. The purpose of the data synthesis is to increase the variation coverage of the head pose. The required training space can be reduced to 2D polar coordinates, i.e., positions of the virtual camera on a viewing sphere around the eye position. During test phase, the input head pose and eye image can be converted into an equivalent 2D polar coordinate representation and a corresponding eye image.

As shown in Fig. 8.7 (left), eye images are synthesized in the range of viewing angles around the eye position where the eye is observable. The range is divided into 6° intervals, and eye images are synthesized at a total of 144 head poses (=virtual camera positions). Figure 8.7 (right) shows examples of the synthesized eye images.

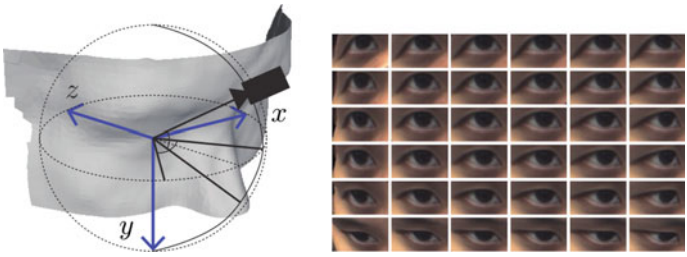

Fig. 8.7 Training data synthesis. *Left* Placement of the virtual cameras. *Right* Examples of synthesized eye images (Reproduced with permission of © 2014 IEEE [45])

Gaze Estimator Training

The training data described above consists of gaze direction vectors, head pose vectors and eye images. The gaze direction is defined as a 2D polar angle vector in the camera coordinate system, and the head pose vector is the rotation vector from the head coordinate system to the world coordinate system. Our goal is to learn a regression function that predicts a 3D gaze direction from the input feature (eye image and head pose).

We use a method based on random forests [3] to learn the regression function. In our problem setting, the input feature consists of multiple modalities, the appearance and the pose of an eye, which are both closely correlated with the output variable, 3D gaze direction. Therefore, we take an approach of learning random forests with some redundancy of head poses.

Figure 8.8 illustrates the structure of our redundant random forests. We cluster training samples into head pose clusters, where each cluster contains samples with

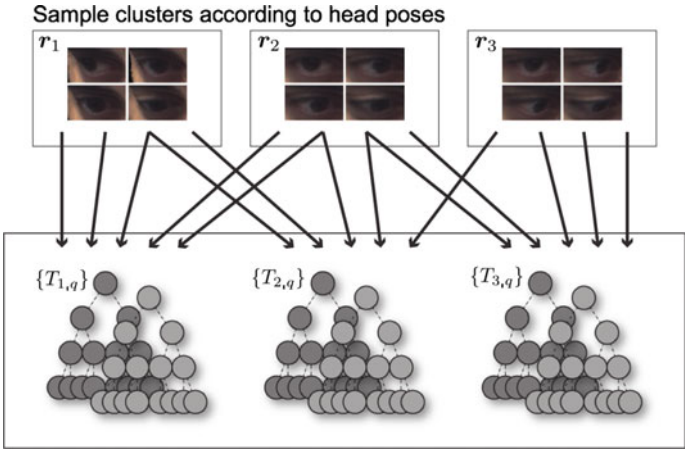

Fig. 8.8 Structure of the gaze estimation function. A set of regression trees with different but overlapping head pose ranges is trained (Reproduced with permission of © 2014 IEEE [45])

the same head pose. Instead of directly learning regression functions for each cluster, we create redundant subsets of the training data to jointly learn regression functions. Namely, to learn a regression function corresponding to the ith head pose, we randomly select training samples from each of the k-nearest sample clusters in the head pose space. A random regression forest is then built using the selected random samples. The test input is queried to its k-nearest regression forests in terms of the distance from the input head pose. Then, the output gaze direction is computed as a mean across all trees of the queried regression forests.

Evaluation

We compared our method with two baseline methods. The first method is the ALR method discussed in Sect. 8.2.1. k-nearest neighbor regression is selected as the second method because of its real-time estimation capability, which is crucial for various gaze applications.

From the dataset, we used 144 synthesized poses for training, and 8 recorded poses for testing. The input image size was set to 15×9 for both training and testing. Figure 8.9 (left) shows the mean estimation errors of all 50 participants for within-subject and cross-subject training. Within-subject errors are evaluated using the target person's own synthesized training data, and this indicates the upper limit of the performance of the proposed learning-by-synthesis approach. Cross-subject errors are evaluated using three-fold cross validation using synthesized training data of 33 different persons. Please note that the cross-subject setting is calibration-free, i.e., not using any training samples from the target person. The proposed method achieved the lowest error with both within-subject and cross-subject training, and the mean error of our method with cross-subject training was $6.5 \pm 1.5°$.

Figure 8.9 (right) shows mean accuracy with respect to the number of training subjects. In the figure, although the accuracy improvement becomes smaller at around 33 subjects, the case of three-fold cross validation discussed above, it does not

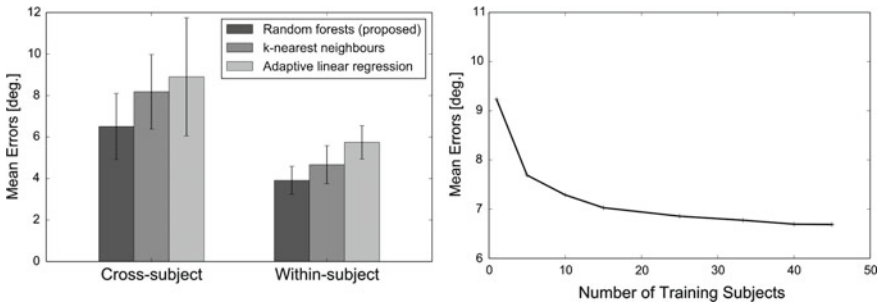

Fig. 8.9 Structure of the gaze estimation function. A set of regression trees with different but overlapping head pose ranges is trained (Reproduced with permission of © 2014 IEEE [45])

apparently converge even with 46 subjects. This result suggests the potential of achieving even greater accuracy by using a larger amount of training data.

8.2.3 Discussion

In Sect. 8.2.1.1, we first introduced a method to solve the core task of appearance-based gaze estimation. The key is the ALR method that optimally selects a sparse set of training samples for gaze estimation. With the proposed method, a gaze tracking system can be implemented that allows for quick calibration.

In Sect. 8.2.2.1, we discussed a method that automatically calibrates a gaze estimation function by using saliency maps. Our method automatically establishes the mapping from the eye image to the gaze point using video clips. Taking a synchronized set of eye images and video frames, our method trains the gaze estimator by regarding the saliency maps as the probabilistic distributions of the gaze points. In our experimental setting with fixed head positions, our method achieves an accuracy with about a 3.5° error.

In Sect. 8.2.2.2, we further discussed a purely learning-based, person- and head pose-independent gaze estimation method. In this method, the gaze estimator is learned using a large amount of synthesized training data. Owing to the synthesized dataset, the learned estimator can estimate gaze directions for arbitrary head poses that are not contained in the original data. The multi-view gaze dataset was made publicly available for future researches.

8.3 Visual Saliency for Subtle Gaze Guidance

8.3.1 Computational Models of Visual Saliency

A large amount of effort for developing computational models of human visual attention has ever been devoted to only *visual* processing. Human visual attention, however, can be easily modulated by other modalities. As an intuitive example, when we hear something interest or strange we tend to look at the direction of sounds even if that direction is not so visually salient. As such, sounds are often strongly related to events that draw human visual attention. We will be able to further augment computational models of human visual attention if we incorporate auditory information into them.

Based on the motivation above, our work [32] proposes a novel model of human visual attention driven by auditory cues. In our model, auditory information plays a supportive role in simulating visual attention, in contrast to standard multi-modal fusion approaches [10, 33, 39, 40]. More concretely, we take an approach that detects visual features in synchronization with surprising auditory events. Our model first

detects *transient* events using the Bayesian surprise model in visual [20] and auditory [41] domains separately, and then looks for visual features in *synchronization* with detected auditory events. Surprise maps are then *modulated* by the selected features.

8.3.1.1 Framework of Our Proposed Model

Figure 8.10 depicts the framework of the proposed model. As shown in this figure, our proposed model consists of four main steps: our model detects transient auditory events, and then selects visual features in synchronization with detected auditory events to modulate the final saliency maps. Note that the proposed model is built on a two-pass algorithm, where the first 3 steps are devoted to selecting visual features that describe major audio-visual events in the input video to produce the final map in the last step.

Bayesian Surprise

The first step extracts surprising events in visual and auditory domains individually where image and audio signals are separately applied to the Bayesian surprise model.

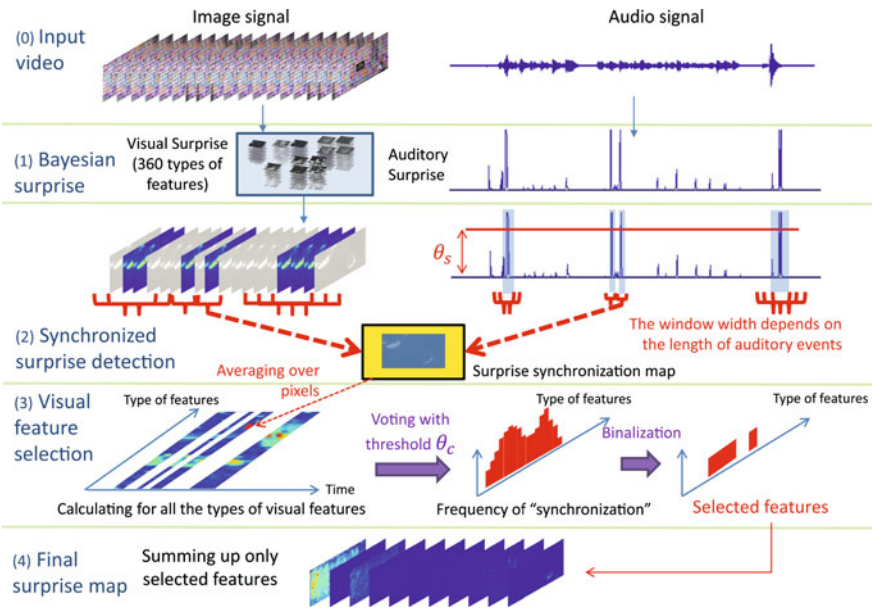

Fig. 8.10 Framework of the proposed model (Reproduced with permission of © 2015 Springer [32].)

For a given input video, 360 visual surprise maps[1] with different types of features and a single auditory surprise signal are extracted.

Synchronization Detector

The second step evaluates synchronization of each visual surprise map with the auditory surprise signal. For this purpose, synchronization detectors are attached to every location in each of the 360 visual surprise maps and the auditory surprise signals, resulting in 360 maps. Every map is averaged over pixels to create a sequence describing how synchronized the corresponding visual surprise map is with the auditory surprise.

A synchronization detector comprises the following three steps. Segments of surprising auditory events are first extracted from the auditory surprise signal using a predefined auditory surprise threshold. For every segment, normalized cross correlation (NCC) is calculated between the auditory surprise signal and visual surprises at every location in each of the 360 visual surprise maps. Every synchronization map is finally averaged over pixels to obtain a sequence that describes how synchronized the visual surprise map is with the auditory surprise.

Features Selection

The third step is devoted to selecting visual features that well synchronize with the auditory surprise. Counting the number of samples with a sufficient level of synchronization for every sequence (we use a predefined correlation threshold here), we obtain a histogram representing the degree of synchronization for every visual surprise map with the auditory surprise. Remembering that every visual surprise map corresponds to a specific type of features, feature selection based on audio-visual synchronization can be implemented by binarizing the histogram, where a threshold for the binarization is adaptively chosen so that its slight change significantly impacts on the number of selected features.

Final Surprise Map

The last step is for forming the final surprise map composed of visual surprise maps with the selected visual features. Only the visual surprise maps of the selected features (with active in the binarized histogram) are accumulated to form the final surprise map. In this way, our proposed model uses a smaller number of visual features than 360 for forming the final map.

[1] 12 feature channels (intensity, 2 color opponents, 4 orientations, temporal onset and 4 directed motion energies) and 6 spatial scales, yielding $12 \times 6 = 72$ feature maps in total. In addition, 5 cascade detectors are implemented at every pixel in every feature map.

Table 8.2 Selected features using the optimal thresholds

	Baseline	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6
Intensity	30	0	8	8	30	0	6
Color	60	4	17	27	60	8	23
Orientation	120	0	46	39	120	0	7
Onset	30	0	0	0	30	1	0
Motion	120	0	0	0	120	14	0
Total	360	4	71	74	360	23	36

8.3.1.2 Evaluation

We experimentally evaluated our proposed model. We selected 6 video clips (advert_bbc4_bees, advert_bbc4_library, sports_kendo, basketball_of_sports, documentary_adrenaline, BBC_wildlife_eagle; we call video 1, 2, etc. in this order), all of which are provided by the DIEM project.[2] We showed them to 15 human subjects. While the subjects were watching the video clips, their eye movements were recorded using an eye tracker Tobii TX300 and then gaze points were detected. As a metric to quantify how well a model predicts actual human eye movements, we used the normalized scan-path saliency (NSS) [20] calculated from the gaze points.

Table 8.2 shows the number of selected visual features by our model with the optimal threshold values for the auditory surprise threshold and the correlation threshold. We can see that only a small fraction of 360 types of features were selected. We also observe that categories such as intensity or color of selected features highly depend on each input video. This is reasonable because what image features are closely correlated with auditory events depends on the video. This also justifies our implementation with a two-pass algorithm.

We compared performances with the state-of-the-art models in addition to the baseline model [20]. They are the saliency map model [21], and the audio-visual attention model using the sound localization [33]. We also compared our model with the model, hereafter called the random feature selection model, in which we randomly selected a given number of image features among all image features used in [20], where the number of features to be selected was set in accordance with the number of image features determined by the optimal threshold values (see Table 8.2).

Figure 8.11 illustrates averages of NSS scores over frames for each video and for each model. We see in Fig. 8.11 that our proposed model produced best NSS scores for all the video, outperforming the other models. Interestingly, the random feature selection model tends to outperform the baseline model. This indicates that using all the image features does not necessarily perform better. Using a smaller number of image features may be better.

[2] http://thediemproject.wordpress.com.

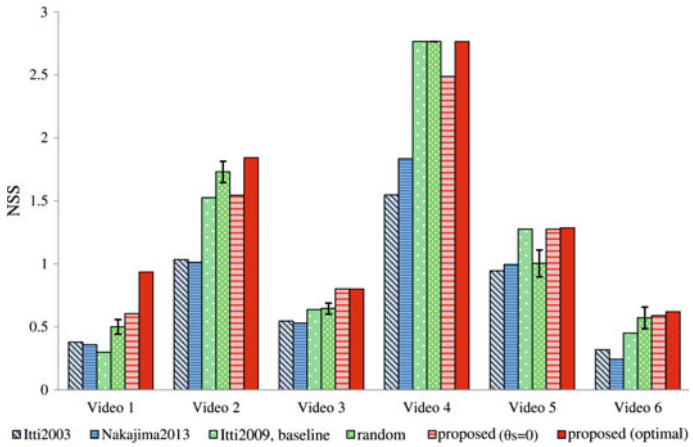

Fig. 8.11 Comparison of NSS averaged over frames for each video (Reproduced with permission of © 2015 Springer [32].)

8.3.1.3 Discussion

Our approach stands on using auditory features as a synchrony cue for selecting visual features. Differently from just fusing audio-visual information, our approach boosts the ability of visual information by selecting visual features synchronized with surprising auditory events.

We used correlation to evaluate synchronization between audio and visual surprises. Mutual information can be also used as an alternative measure for synchronization [38]. We can improve our model into several directions in future, e.g. the introduction to adaptive image feature selection depending on the auditory event or the location in the image and machine learning strategies for capturing generic structures of audio-visual events.

8.3.2 Saliency-Based Gaze Guidance

With the rapid progress of computer facilities, computer usage in every aspect of our daily life has become more and more popular. In fact, we have a drastically increased number of information systems such as the smartphone or the tablet PC. Moreover, electronic bulletin boards can be found everywhere and the electrical audio assistant is available in many cities. These information systems should be in good harmony with human beings and, thus, developing a technology for naturally guiding the human gaze is a key issue.

In order to develop such a technology, we put ourselves in the position of using visual saliency and directly modulate saliency for guiding human visual attention.

Namely, our work [13] proposes a method for modulating an image, for a given region in the image, to synthesize the image in which the region shows the highest saliency. Our method iteratively modulates the intensity and color so that the saliency inside the region increases while that outside the region decreases. This iteration is carried out until we obtain the image in which a given region is most salient over the entire image. With the image obtained in this way, we can smoothly attract human visual attention to the given region, without any interruption to the human gaze.

8.3.2.1 Saliency-Based Image Modulation

We assume that for an input image, we are given a region that should draw human visual attention. A given region is not always salient in an input image. We thus iteratively modulate the image so that the given region becomes most salient in the entire image. We note that our method is restricted to modulating only the intensity and color.

Control Saliency

In order to raise the saliency of a given region, we have to strengthen visual features inside the region. Furthermore, weakening the visual features outside the region also contributes to reducing the saliency outside the region; as a result, the saliency inside the region increases. The visual features mentioned here refer to the intensity and color.

Controlling the visual features depending on the location in the image is more effective than uniformly controlling them over the entire image. This is because saliency at a pixel is computed based on how discriminative the pixel is from its surroundings. Therefore, the relationship in visual features between a pixel of interest and its surroundings is a key in controlling saliency.

Image Modulation

We introduce two of parameters in our modulation: the saliency-based weight w_p and the ratio for saliency value $Q_p = (Q_p^{\mathrm{R}}, Q_p^{\mathrm{G}}, Q_p^{\mathrm{B}})$ where p denotes a pixel and R, G, B indicate red, green, blue channels respectively. The saliency-based weight adjusts the direction of change when taking into account whether the pixel of interest is inside or outside the region. The ratio for saliency value, on the other hand, determines the degree of change of visual features when taking into account difference from the surrounding area. We remark that the saliency-based weight depends on the pixel position while the ratio for saliency value depends on both the pixel position and RGB-channels.

At each iteration, RGB-values at each pixel are updated:

$$R'_p = R_p + w_p Q_p^{\mathrm{R}},$$
$$G'_p = G_p + w_p Q_p^{\mathrm{G}},$$
$$B'_p = B_p + w_p Q_p^{\mathrm{B}},$$

where R_p, G_p, B_p respectively denote red-, green-, blue-channel values of pixel p before the updating, and R'_p, G'_p, B'_p denote those after the updating.

The above updating may cause significant difference of the updated image in visual perception from the input image. To reduce such difference as much as possible, we introduce adjustment into RGB-channel values. Namely, for each channel, we introduce a bias and dynamic range adaptation so that the histogram of channel values over the image after the updating becomes as similar as possible to that before the updating. This can be formulated as the weighted least squares method where weights are determined by taking into account the saliency-based weight.

Using the obtained bias $\alpha = (\alpha^{\mathrm{R}}, \alpha^{\mathrm{G}}, \alpha^{\mathrm{B}})$ and dynamic range adaptation $\beta = (\beta^{\mathrm{R}}, \beta^{\mathrm{G}}, \beta^{\mathrm{B}})$, we adjust the updated image as follows.

$$\widetilde{R'}_p = \beta^{\mathrm{R}} R'_p + \alpha^{\mathrm{R}},$$
$$\widetilde{G'}_p = \beta^{\mathrm{G}} G'_p + \alpha^{\mathrm{G}},$$
$$\widetilde{B'}_p = \beta^{\mathrm{B}} B'_p + \alpha^{\mathrm{B}}.$$

We iterate the above pair of updating and adjustment until the given region becomes most salient in the entire image. In this way, our modulated image guarantees that a given region is most salient while minimizing visual perceptual difference of the modulated image from the input image.

Saliency-Based Weight

The saliency-based weight depends on not only the saliency value of a concerned pixel but also whether or not the pixel is inside the region. The saliency-based weight of a pixel inside the region should be positive while that outside the region should be negative. This is because we increase the saliency inside the region and decrease the saliency outside the region. The magnitude of the saliency-based weight of each pixel may be proportional to the saliency value of the pixel. We, however, smooth it over the entire image in order to avoid drastic impact caused by change in sign at the region boundary. Therefore, after attaching an appropriate sign to the saliency value of each pixel, we apply the Gaussian filter to have the saliency-based weight at the pixel.

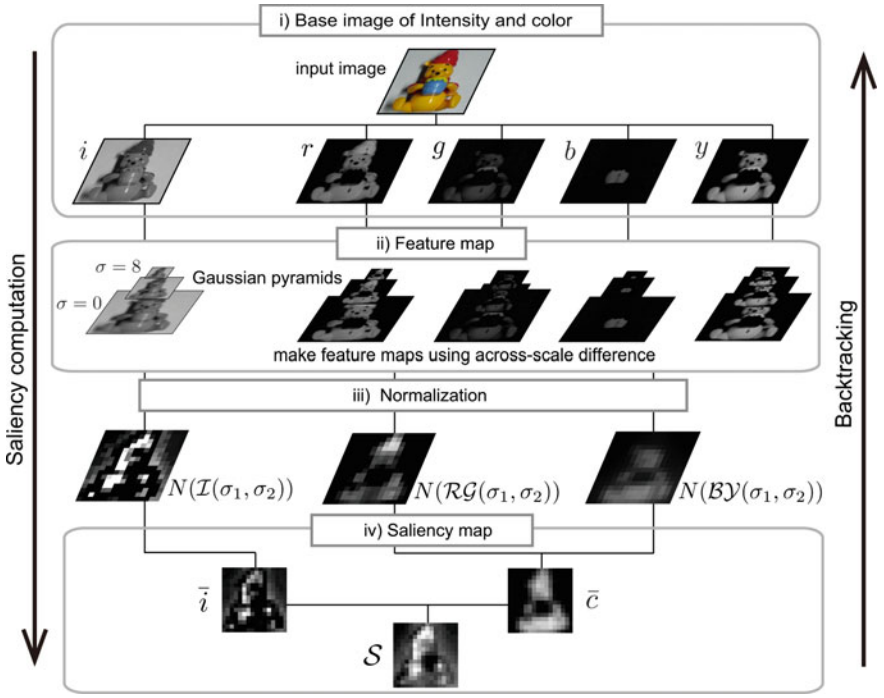

Fig. 8.12 Saliency computation and its backtracking

Ratio for Saliency Value

The ratio for saliency value reflects how much each feature influences the saliency of a concerned pixel. For example, if the red color heavily influences the saliency, we enlarge the red-channel ratio for saliency value (i.e., Q_p^{R}) of the pixel. We thus evaluate which color or intensity provides great impact on the saliency of a pixel. To compute the influence, we return to the procedures for computing the saliency map. We can identify how each visual feature influences on the saliency at a given pixel by backtracking one by one through the saliency-computation procedures (Fig. 8.12).

8.3.2.2 Evaluation

We experimentally evaluated our proposed modulation method. We prepared 40 different color input images of 512 × 256 pixels. We specify a region to each input image. Furthermore, we selected 10 input images and specify another region to each so that we have 50 modulated images in total. We note that we chose a less salient region in an input image to specify a region. For a pair of an input image and a specified region, we applied our modulation method. We also applied our method

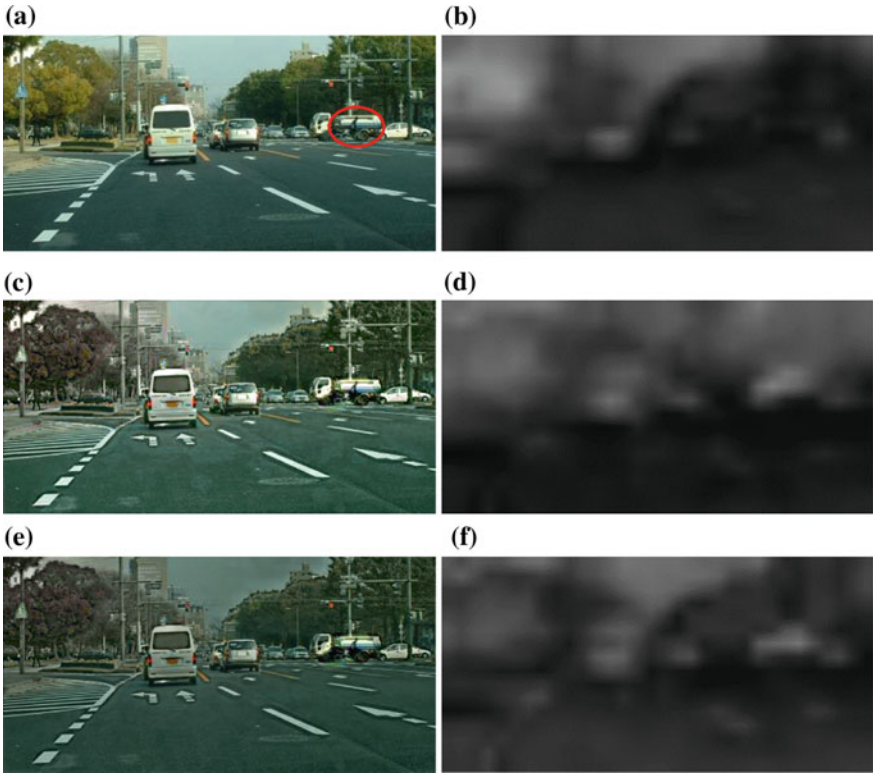

Fig. 8.13 Examples of input image and modulated images with and without adjustment. The *red circle* in (**a**) is a specified region to enhance saliency. **a** Input image. **b** Saliency map of input image. **c** Modulated image with adjustment. **d** Saliency map of the modulated image with adjustment. **e** Modulated image w/o adjustment. **f** Saliency map of the modulated image w/o adjustment

without adjustment (namely, without bias and dynamic range adaptation) to see how our adjustment is effective.

Figure 8.13 illustrates an example of our input images and modulated images together with corresponding saliency maps. As we see, our specified region becomes most salient over the image after the modulation. We can also observe that our adjustment contributes to keep visual perceptual similarity of the modulated image to the input image.

We also evaluated how modulated images draw gaze points to the specified regions. We randomly showed 90 images (40 input images and 50 modulated images) to 24 human subjects where each image was shown for three seconds. We remark that for the gaze point initialization, we displayed a white image with the cross located in the center between successive two images. While the subjects were looking at the images, their eye movements were recorded using an eye tracker Tobii TX300 and then gaze points were detected. We then evaluated whether or not the specified region of each image drew the gaze points.

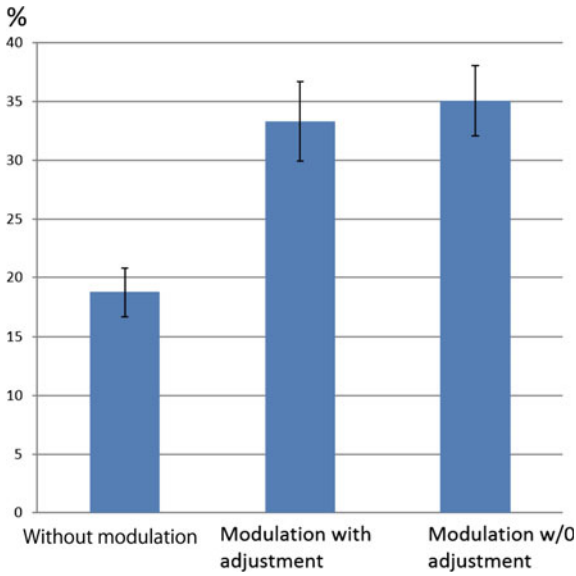

Fig. 8.14 Average rates for drawing visual attention

Figure 8.14 shows the average of ratios for drawing gaze points to a specified region. We see that modulated images significantly improved the ratio. Indeed, our hypothesis testing statistically confirms that the ratios before and after the modulation are significantly different from each other.

8.3.2.3 Discussion

We discussed an image modulation method for naturally guiding human visual attention. Our method is different from conventional methods [1, 26] in that our method is based on a fully bottom-up information gathering obtained from an image. Namely, we stood on the position of using visual saliency and presented iteratively modulating the intensity and color of an image until the saliency of a given region becomes highest over the entire image. Experimental results confirm that our image modulation method draws human visual attention toward our specified region.

The advantage of our method is that we do not need to present any visual stimulus to a subject in order to attract his visual attention. Incorporating human perceptual properties into our image modulation is left for future work.

8.3.3 Gaze Guidance for Interactive Systems

In most interactive systems, it is often the case that information contents providers want to guide viewers' attention to a particular location of the display. For example in

digital signage, the contents providers want people to look at their products rather than models who are smiling on the display. On web pages, they want to guide people's attention to banner advertisements. In electronic markets, people's attention should be guided to today's campaign products.

There have been some approaches which guide people's attention to a particular location of the display. These approaches include changing the color of the object, flashing the object, vibrating the object, and so on to draw people's attention. Such "active" methods, however, are not appreciated by the people. For example, flashing or animation would interrupt the people's concentration to their main task. The web pages using such visual effect would become unpopular.

On the other hand, there are some studies to guide people's attention without making much modification to the display. Some of them can even move their focus without being recognized the modification by focusing on the characteristics of human's visual perception. Recent work [1, 14] uses saliency map by Itti et al. [22]. However, the approaches using the saliency map often changes the color and intensity of the image, and therefore the modified image often become unnatural.

We proposes a method to guide people's visual attention to the intended location without being recognized by the people using dynamic resolution control.

8.3.3.1 Visual Guidance Using Dynamic Resolution Control

The human visual system is always searching for something in the real world. When it finds something interesting, it would stay on the object a little longer (i.e. fixation) and then keep on moving. When the image is blurred, although it depends on the blur strength, the human visual system cannot obtain much information and it would change its focus to other objects. However, if it finds the object with high resolution, the visual system would stay on the object. Our attention control mechanism uses such characteristics of the human visual system. The whole image is blurred while the region to which we want to guide people's gaze is remained in high resolution.

By using this approach, we could move people's focus to our intended position. However, one problem of this approach is that people easily recognize the attention control when we use the strong blur. We do not want people to realize the blur.

Our solution to this problem is that we first show the original high resolution image to the people. Then the image is gradually blurred until they realize the blur. When the people's gaze is guided, the image is gradually recovered to its original resolution. This process is illustrated in Fig. 8.15.

In our experiment, the Gaussian filter was used. The Gaussian filter is one of the smoothing filters and it is calculated as $f(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$, where $x = y = 13$ in our study.

In order to confirm the effectiveness of our approach, it is necessary to show the following.

- Is it possible to move people's attention by using resolution control?
- Is there any threshold in blur level at which people are aware of the blur?

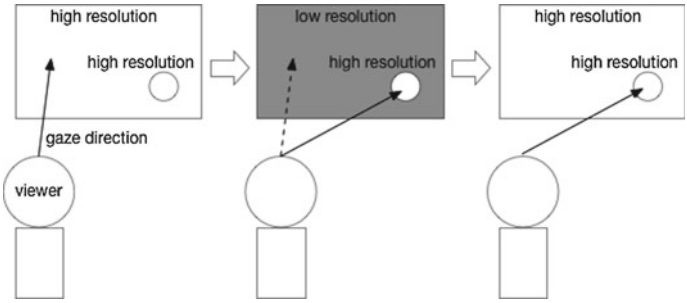

Fig. 8.15 A basic concept of navigating users' attention using resolution control

In the following sections, we describe two experiments which are conducted to answer the above questions.

8.3.3.2 Experiment 1

Figure 8.16 shows the set up of this experiment. There is a 23 inch LCD display (1920 × 1080 pixel) on the table with a Tobii TX300 eye tracker. In order to measure accurate gaze direction, a chin rest was used. The distance between the eye and the display is 60 cm. Subjects were allowed to see the image on the display freely.

Figure 8.17 shows the results of the experiment. By seeing the heat maps, it is understandable that users' gaze were statistically guided on the high resolution region. We did not tell the subjects before the experiment that there is a high resolution area in the image. When the experiments finished, the subjects were asked if they recognize that there is a high resolution region. They, however, answered that they did not recognize it.

Figure 8.18 (left) is a graph which shows a relation between the blur strength (σ) and the time to take until subjects' gaze first enter to the high resolution area. A horizontal axis shows the blur strength and a vertical axis shows the time. When $\sigma = 1$, it took 5.5 s in average. However, when $\sigma \geq 2$, it took 3 s in average. It is clear that the subjects' gaze was guided successfully.

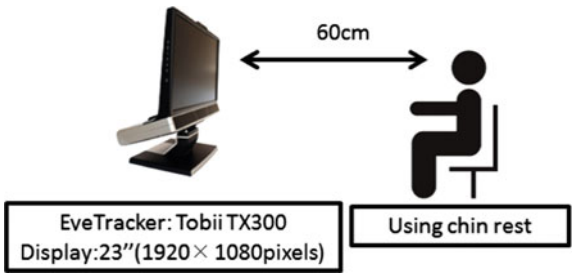

Fig. 8.16 A set up of the experiments (Reproduced with permission of © 2015 IPSJ [15])

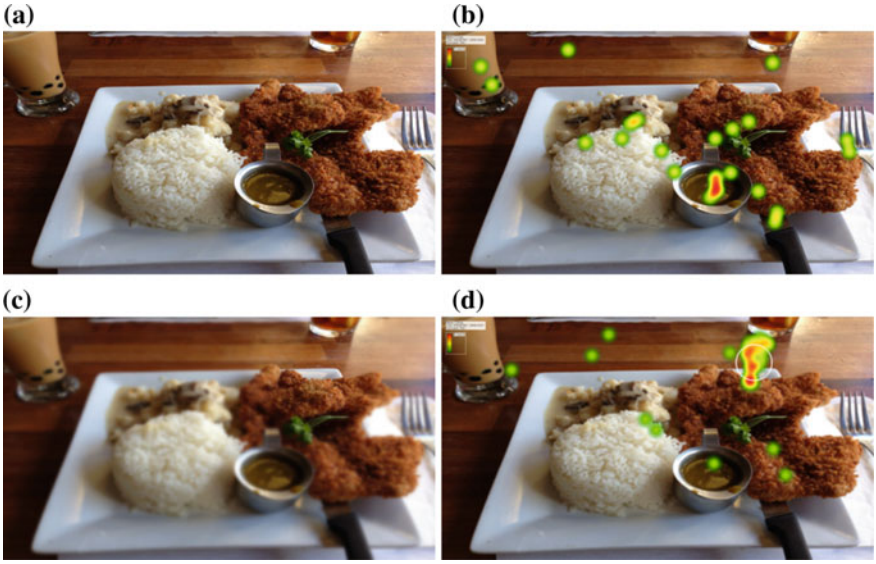

Fig. 8.17 The images used in the first experiment and their heat maps (Reproduced with permission of © 2015 IPSJ [15]). **a** A presented image when $\sigma = 0$. **b** A heat map when $\sigma = 0$. **c** A blurred image when $\sigma = 5$. **d** A heat map when $\sigma = 5$. A white circle indicates the high resolution region

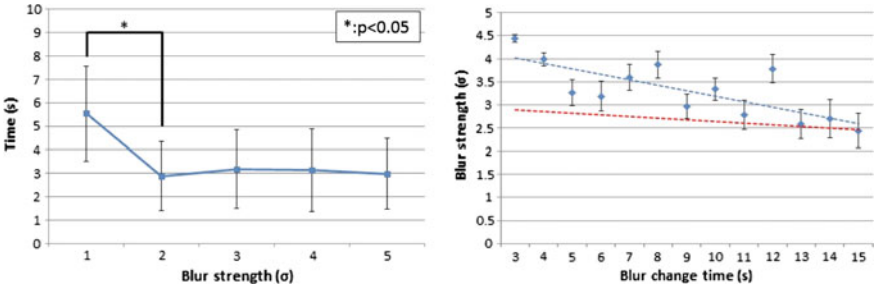

Fig. 8.18 Results of the experiment 1 (*left*) and experiment 2 (*right*)

8.3.3.3 Experiment 2

The purpose of this experiment is to investigate when people are aware of the guidance. In other words, we want to know σ at which people are aware of the resolution control. And if that $\sigma > 2$, it is said that we can guide people's gaze without being recognized.

We used the same hardware set up for the experiment as the experiment 1. The image was shown on the display and was gradually blurred from $\sigma = 0$ to $\sigma = 5$ within different transition time (3–15 s). Subjects were asked to click the mouse when they recognized the blur.

Figure 8.18 (right) shows the result of the experiment 2. This graph shows the relation between the time to be completely blurred (i.e. $\sigma = 5$) and the blur strength when the subjects clicked the mouse button. From this graph, it is understandable that there is a certain threshold at which the subjects recognize the blur. And the threshold depends on the transition time. The graph shows that the subjects were easier to recognize the blur if the transition time was long and were more difficult if the transition time was short. This is because if the transition time is long the subjects have enough time to find the area which is easier to recognize the blur.

8.3.3.4 Discussion

One of the advantages of our approach is that it can be used in multiple person's environment. Bailey's approach [1] tracks one person's gaze and make a flash while he/her is not seeing the intended position. In multiple person's environment, someone may see the flash. On the other hand, our approach can be applied to multiple person's environment. The resolution control is done below a threshold that people can recognize the blur. Even if multiple people see the different location simultaneously, it is hard to recognize the blur.

Subliminal stimuli [24, 43] are stimuli below a threshold for conscious perception. Its effectiveness has been under discussion in cognitive studies for long years. However, since there is a concern that it may force to embed information to people's brain without being recognized, its use has not been allowed on TV or movies. Since our method guides people's attention without being recognized, there might be the similar concern that our method is a kind of subliminal. The main issue of the subliminal is that it tries to send information which cannot be recognized by the user. On the other hand, our approach does not hide any information even though it navigates their attention without being recognized. At this point, our approach is essentially different with subliminal.

Here is the summary.

- People's attention is guided to the high resolution area.
- The effect of gaze guidance is obvious at $\sigma \geq 2$.
- People are not aware of the blur below a certain threshold.
- The longer the transition time is, the smaller the blur strength at which people become aware of the blur.

From above, we conclude that our method can guide people's visual attention without being noticed.

8.4 Human Gaze Control by Robots

Suppose we have the following situation: You would like to show a prototype of your new product to a colleague. He is working in an office shared by several other people. You enter the office and approach to his desk. You notice him reading some

report. You do not want to disturb him, so you wait until he lifts his face from the report. Taking this opportunity, you look at his face, making eye contact. Then, you may say, "Hello," and turn your head to look at the prototype in your hand. He also looks at it. You turn your head toward his face again, saying, "What do you think?" He knows what you are talking about. Here, by looking at your colleague using the proper timing, you can attract his gaze and attention toward you, and by eye contact, you establish a communication channel with him. After that, you can expect to have established mutual gaze patterns with him. If you look at anything, he surely looks at the same object. In other words, you can control his gaze.

In this project, we propose a robot that can initiate interactions with a human in a socially acceptable manner, as described above. Most human-robot interaction studies consider cases where either robots and humans are already interacting, or where humans who require the robots' services call the robots by voice and/or nonverbal behaviors such as hand gestures to initiate interaction with the robots. However, there are cases where a robot needs to initiate interaction with a human being. The robot also needs to behave in a socially acceptable manner in such cases. To do so, we propose to consider the human's level of visual focus of attention. The Visual Focus Of Attention (VFOA) is the behavioral and cognitive process that indicates where and at what a person is looking, and in computer vision, it is mainly determined by eye gaze and head pose dynamics [42]. The Level of Visual Focus Of Attention (LVFOA) refers to how much concentration is given to a particular VFOA and is classified into discrete levels: low, high, or medium [8]. If the robot needs to start communication urgently such as during an emergency, it does not need to consider the current situation of the person. Otherwise, the robot should observe the person to know at what/who s/he is looking (VFOA) and how attentively s/he is doing so (LVFOA). Then, it should determine the proper timing to attract her/his attention so that it does not interfere with her/his current activities such as work. We propose a system in which the robot interacts with the target person intelligently and in a socially acceptable manner so that it can interact by considering her/his current VFOA as well as other people in the environment.

In Sect. 8.4.1, we describe how the robot can control a target person's gaze to attract her/his attention and establish mutual gaze based on the level of visual focus of attention (LVFOA). In Sect. 8.4.2, we describe our robot head with eyes designed for effective gaze communication as described in Sect. 8.4.1.

8.4.1 Initiating Interaction from a Robot Based on the Level of Visual Focus of Attention

The VFOA is an important cue for attracting attention and initiating interaction because—(i) it helps with understanding what the person is doing and (ii) it indicates addressee-hood (who is looking at whom). For instance, if the target person's VFOA is toward the robot, the robot can immediately establish a communication channel

through eye contact. If the target person is involved in some task, the robot should wait to find the proper timing to attract her/his attention and establish a communication channel. In this research, the proper timing is determined by detecting the level of attention of the target person on her/his current task. In a scenario such as reading, writing, or browsing, the robot should initiate interaction with the target person when her/his level of attention is low. In other settings, such as at a museum, the robot may need to consider people's high level of attention towards exhibits to provide guidance on objects of interest.

8.4.1.1 Proposed Approach

The proposed approach is illustrated in Fig. 8.19a, b. In the *initiating interaction module* (Fig. 8.19a left), the robot recognizes and tracks the target person's VFOA. If they are initially face-to-face, the robot generates an awareness signal and makes eye contact with the target person. Otherwise, the robot tries to attract the target person's attention by recognizing her/his current task. The robot detects the level of the current VFOA until T_s (where T_s is the maximum span of sustained VFOA, which is explained in the next section). The robot uses either a low or high level of current VFOA (depending on the person's current task) at time t as its trigger to generate an attention attraction (AA) signal (weak or strong) depending on the viewing situation of her/his shifted VFOA. A person's field of view is divided into central and peripheral visions. We represent the viewing situation (relation between the target person's gaze (face) direction and the robot position) by where the robot is seen in the field of view of the target person. We classify it into the three regions: Central Field of View (CFV), Near Peripheral Field of View (NPFV (RNPFV: the right side, and LNPFV: the left side)), and Far Peripheral Field of View (FPFV (RFPFV: the right side, and LFPFV: the left side)) [18, 19, 50].

If the VFOA is detected in the CFV/LNPFV/RNPFV, then the robot generates a head turning action (weak signal). However, if the detected VFOA is in the LFPFV or RFPFV, then the robot generates a head shaking action (strong signal). Figrue 8.20 illustrates these classified regions when the camera is placed in the CFV region. We define the angular regions based on the detected frontal and profile faces. For example, in the CFV region, we detect frontal faces only. However, in the other regions, we may detect two face patterns such as a half-pose right profile face and a full-pose right profile face in the LFPFV region.

Once the robot succeeds in attracting the target person's attention toward it, the *communication channel establishment module* (right part of Fig. 8.19a) tries to establish a communication channel with her/him. For this purpose, the robot determines the level of shifted attention toward it. Based on the level of shifted attention, the robot generates an awareness signal toward the target person to indicate that it wants to communicate with her/him. Finally, the robot makes eye contact through eye blinking to establish a communication channel.

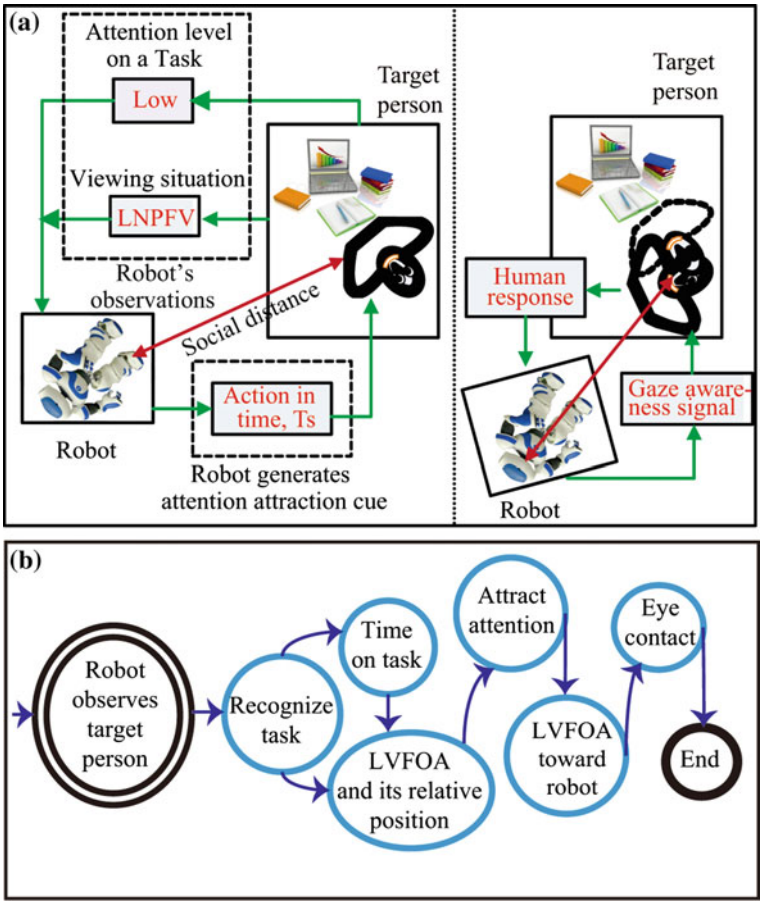

Fig. 8.19 **a** An abstract view of the proposed approach, and **b** basic steps of the proposed approach (Reproduced with permission of © 2015 IEEE [9])

8.4.1.2 Recognition of VFOA and Its Level

We are interested in detecting: *sustained attention* and *focused or shifted attention*. Focused or shifted attention is a short-term response to a stimulus or any other unexpected occurrence. The span or length of this attention is very brief [6] and after a few seconds, it is likely that the person will look away, return to the previous task, or think about something else. Sustained attention on the other hand is the level of attention that produces consistent results on a task over time. The duration of sustained attention also depends on the task. To learn about this, we observed humans working on various tasks and measured the duration. For our system, we use the maximum value for each task obtained in the observations as the maximum waiting

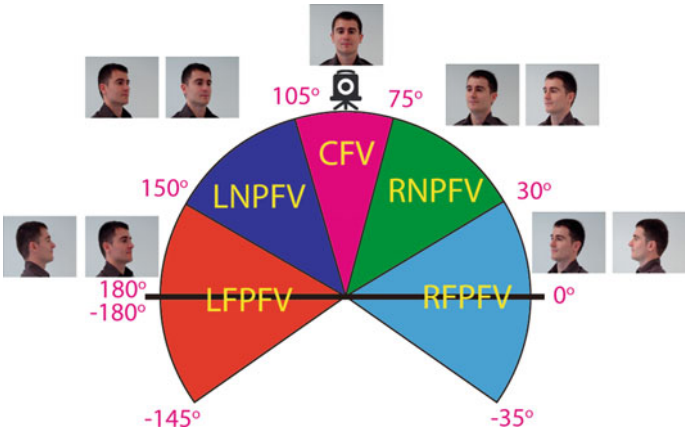

Fig. 8.20 Classification of head orientation into five angular regions. The faces shown from the GTAV face database [48] (Reproduced with permission of © 2015 IEEE [9])

time for the robot when it cannot find the proper timing for initiating interaction. We also use visual cues, gaze pattern, and task context to recognize VFOA and estimate its level.

Visual Cues

We detect and track the head to obtain the head pose. We also detect head movements, in particular, to find head movements toward the robot.

Gaze Pattern

A person's gaze pattern indicates her/his object of interest [12]. In general, human gaze patterns are classified into three viewing categories, distinguished by context: *spontaneous viewing, task or scene-relevant viewing,* and *orientation of thought viewing* [25]. *Spontaneous viewing* occurs when a person views the scene without any specific task in mind, i.e., when s/he is "just seeing" the scene. *Task or scene-relevant viewing* occurs when a person views the scene with a particular question or task in mind (e.g., s/he may be interested in a particular painting in the museum). *Orientation of thought viewing* occurs when the subject is not paying much attention to where she is looking, but is attending to some "inner thought". We consider the former two. In this research, we consider a gaze pattern as that which is constructed by considering the effects of both head movements and eye gaze. We classify gaze patterns using the support vector machine classifier [23]. Figure 8.21 shows examples of gaze pattern recognition.

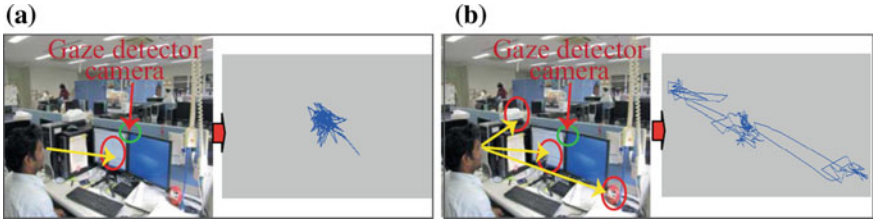

Fig. 8.21 Gaze pattern: **a** task or scene-relevant viewing, **b** spontaneous viewing (Reproduced with permission of © 2015 IEEE [9])

Task Context

Task context is determined by recognizing the task in which the target person is involved. For instance, if the target person is involved in a "reading" task, then the contextual cue such as "downward head" indicates that her/his attention is toward the book. However, the "page turn over", or "upward the head" behaviors indicate that the person loses her/his attention. We consider four tasks in this project: reading, writing, browsing, and viewing painting. We use the histogram of orientation gradient (HOG) feature [7] to train the SVM classifier to recognize tasks. For each task, we use the task related VFOA span, (T_s) to determine how long the robot should wait or within which period of time the robot interacts with the target person. We also define some task specific cues to determine the level of attention. For example, in the reading situation, we use the "page turn over" behavior and deviation in tilt angle cues to measure the LVFOA.

LVFOA Recognition

The level of VFOA is classified into two categories (low or high) based on the contextual cues, and gaze pattern. When the level of attention goes low, the system assumes that a loss of VFOA is detected. In any case, if spontaneous viewing is detected, then it is assumed that the person has no particular attention on a task. Thus, a low attention level is detected. In addition, we use the context cues for the current task. For example, in the case of reading and writing tasks, in addition to head pose changes, we also consider the "page turn over" and "stop writing" behaviors for detection of low attention level.

We tested the robot system in an office scenario and a museum scenario and confirmed that it works as expected. We compared our robot with one that does not consider the contextual situations of people. We found that our robot obtained a favorable impression from the participants. Details are found in [9].

8.4.2 Design of a Robot Head for Gaze Communication

Human eyes not only serve the function of enabling us "to see" something, but also performs the vital role of allowing us "to show" our gaze for non-verbal communication, such as through establishing eye contact and joint attention. The eyes of service robots should therefore also perform both of these functions. Moreover, they should be friendly in appearance so that humans would feel comfortable with the robots. Therefore we maintain that it is important to consider the capacity for gaze communication and friendliness in designing the appearance of robot eyes. In this project, we examined which shape for robot eyes is most suitable for gaze reading and gives the friendliest impression, through experiments where we altered the shape and iris size of robot eyes.

8.4.2.1 Eyes for Accurate Gaze Reading

Eyes and gaze have been examined in various fields. In biology, Kobayashi and Kohshima [27] found that among primates, only human eyes have no pigment in the sclera; moreover they also have the horizontally longest shape with the largest exposed area of sclera. Various explanations were offered as to why other primates have sclera colored in a similar fashion to their irises or the outside of their eyes. But all the explanations were based on the consensus that primates might be avoiding clearly showing their gaze. In contrast, human eyes have sclera with clearly different colors from those of the irises and the outside of the eyes. This enables human gaze to be readily comprehended by others. Kobayashi and Kohshima proposed the hypothesis of "gaze grooming" as the reason for why human eyes have this feature. From this study, we decided to focus in our own study, on whether the difference in the shape of the eyes, changes the ease of gaze reading.

Based on the eye parameters of pigments focused on by Kobayashi and Koshima [27], we changed the lid distance of the eyes when preparing design candidates. We prepared three types of outline shape for the eyes, specifically "round," "ellipse," and "squint." These three shapes were generated by setting the lid distance at 1.0, 0.5, and 0.25 times as long as the eye width, respectively. In the same manner, we changed the iris diameter to 0.75, 0.5, and 0.25 times as long as the eye width, and labeled them "large," "medium," and "small," respectively. Consequently, we had 9 types (the combination of 3 outline shapes and 3 iris sizes) of eye design, as shown in Fig. 8.22. A spherical shape and medium gray color were employed for the robot face in all instances to negate any effect of facial design. Notably, the eye employing the ellipse outline shape and the medium iris diameter (Fig. 8.22-E) is the most similar to the human eye in terms of the ratio of these parameters.

We developed a robot head as shown in Fig. 8.23 to examine which shape of robot eyes is most suitable for gaze reading. Each of the eyes consists of a projector, a mirror and a screen. The eye images, generated by CG, are projected onto the hemisphere screen via rear-projection. By using this projection mechanism, we can change the

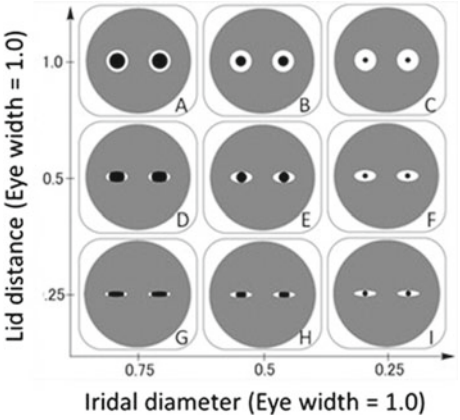

Fig. 8.22 Candidate designs for robot eyes derived by varying lid distance and iridal diameter (Reproduced with permission of © 2013 John Benjamins Publishing Company [36].)

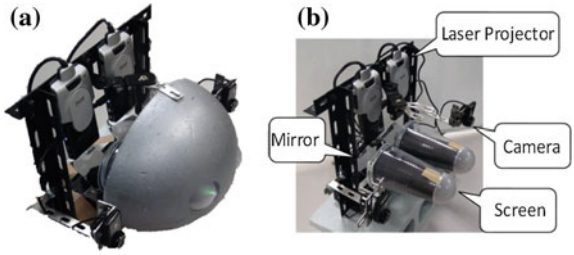

Fig. 8.23 **a** Overview of our proposed robot head. **b** The inside of the robot head consists of laser projectors, mirrors and screens (Reproduced with permission of © 2013 John Benjamins Publishing Company [36].)

iris diameter. In addition, by replacing the mask, we can change the outline shape of the eyes.

We then conducted experiments to analyze the relationship between the accuracy of gaze reading and the shape of the robot eyes by using this robot head. We evaluated errors in gaze reading using the nine types of design candidates for robot eyes shown in Fig. 8.22. In the experiments, we lined up a series of markers between the participant and the robot head, as shown in Fig. 8.24. We asked the participant to stand in front of the robot head, face-to-face, with his/her head fixed on the mount. We then asked the participant to state at which marker the robot looked. Figure 8.25 shows the result. See details of the experiments in [36].

8.4.2.2 Friendly Eyes

We sought to examine the impression of friendliness by changing the outline shape of the eyes and the size of the iris, and seeing how participants responded. We conducted experiments to ascertain the apparent friendliness of the nine types of robot eyes shown in Fig. 8.22. We evaluated the degree of apparent friendliness by using Thurstone's method of paired comparison. We developed a web-based system

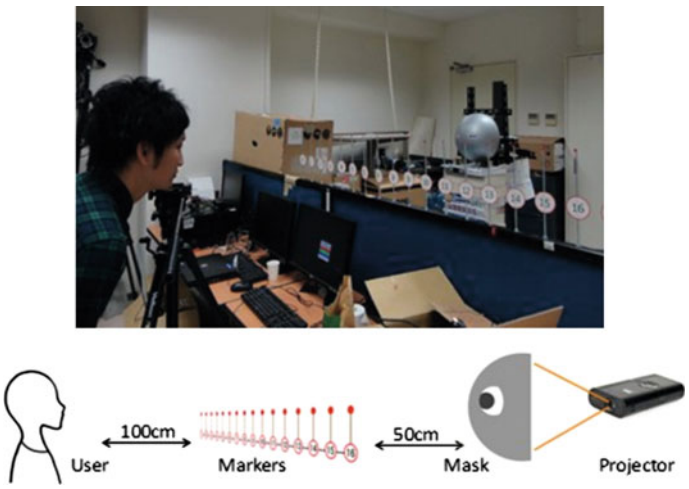

Fig. 8.24 Experimental scene of robot gaze reading (Reproduced with permission of © 2013 John Benjamins Publishing Company [36].)

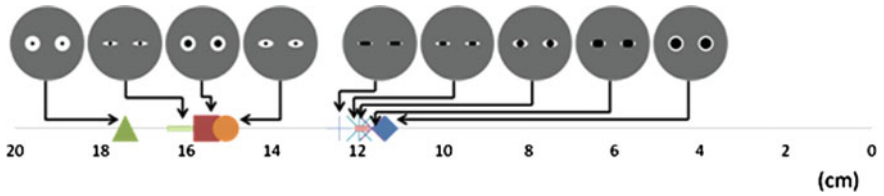

Fig. 8.25 Errors of gaze reading for each design candidate for robot eyes. The smaller value indicate smaller error and thus a more readable gaze (Reproduced with permission of © 2013 John Benjamins Publishing Company [36].)

for collecting answers from participants. Participants could choose one of the two images of a pair of robot eyes by tapping the iPad screen. We asked participants to answer the question, "These are robot faces. Which face do you think is friendlier?" for all 36 pairs of combinations of the nine types of robot eyes, which appeared in random order. We note that in actuality, the question was in Japanese. We used 105 participants: 60 males, 43 females, and 2 no-records. They were Japanese students of the school of liberal arts at Saitama University. We then analyzed the result by using Thurstone's method of paired comparison for scaling the impression of the robot eyes. Figure 8.26 shows the result.

8.4.2.3 Design Principles of Robot Eyes

From these experimental results, we have established the design principles of the robot eyes, and have developed a robot head based on the principles. We had expected

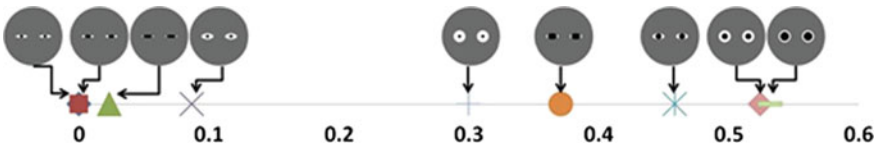

Fig. 8.26 Result of scaled experiment in impression of robot eyes by Thurstone's method of paired comparison, case III model by Japanese students. The graph was made by scaling it so that for Fig. 8.22-I, smallest value is 0. The more chosen a shape of eyes is by participants, the higher the scale value that shape of eyes gains (Reproduced with permission of © 2013 John Benjamins Publishing Company [36].)

that the human-like eye (F) would show the most accurate result in gaze reading as Kobayashi and Koshiba indicated [27]. However, the design with a round outline shape and a large iris (A) gave the best result. Although the differences from the most human-like eyes (E), D, H, and G are small. The results are not good if the iris is small compared to the eye's size (F, B, I, and C). Therefore, we can conclude that the iris should be large compared to the eye's size for accurate gaze reading. The impression of eyes may also depend on nationality, race, and other factors. Our participants (Japanese students) most preferred the design with a round outline shape and a large iris (Fig. 8.22-A). In addition to the experiments described above, we examined which head shape: a sphere, a flat plane, or a sphere with a nose could give the most accurate gaze reading result by a similar experiment described in Sect. 8.4.2.1 [35]. From these, we can conclude that a robot face with the eyes of a round outline shape and a large iris (Fig. 8.22-A) and with a nose is most suitable for gaze reading and conveying an impression of friendliness. We have developed such a robot head as shown in Fig. 8.27. This research provides the basic principles of robot eye design although we still need to consider various other factors in designing robot heads.

Fig. 8.27 Developed robot head with projection eyes

8.5 Conclusion

In this chapter, we presented techniques developed in our CREST project for sensing and guiding our gaze without distracting our activities. For remote gaze sensing with less or no calibration effort, we introduced three key ideas. Firstly, we proposed an appearance-based gaze sensing method with adaptive linear regression (ALR) that optimally selects a sparse set of training samples for gaze estimation. The method achieves higher accuracy of gaze estimation with significantly fewer training samples of low resolution eye images than existing appearance-based gaze estimation methods. Secondly, we exploited the new approach of carrying out auto calibration of gaze sensing from user fs natural viewing behavior predicted with a computational model of visual saliency. Lastly, we introduced a user-independent single-shot gaze estimation method. The key idea is to learn a generic gaze estimator by using a large dataset of eye images collected for different people, head poses, and gaze directions.

For guiding human gaze to desired locations in a non-disturbing way, we studied two approaches for gaze control. The first approach is subtle modulation of visual stimuli based on visual saliency models. We have shown that our gaze can be guided to a desired region in the visual stimuli (a given image) by modulating intensity or color contrast of the region, to the level just enough to make the region stand out. Related to the gaze guidance based on visual saliency, we also studied a new approach of enhancing visual saliency by integrating inputs from different modalities, more specifically, augmenting a visual saliency model by incorporating auditory information. The second approach for gaze guidance is to control human gaze by using robot's non-verbal behavior in human-robot interaction. To allow a robot to initiate interactions with a human in a socially acceptable manner, we introduced a model of human-robot interaction based on the level of visual focus of attention of a user. We also presented design principles of robot eyes in both their appearance and motion. Experiments carried out by using a prototype robot demonstrated the effectiveness of the proposed model of interaction and the eye design.

Acknowledgments The work presented in this chapter was supported by CREST, JST.

References

1. R. Bailey, A. McNamara, N. Sudarsanam, C. Grimm, Subtle gaze direction. ACM Trans. Graph. (TOG) **28**(4), 100 (2009)
2. A. Borji, L. Itti, State-of-the-art in visual attention modeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 185–207 (2013)
3. L. Breiman, Random forests. Mach. Learn. **45**(1), 5–32 (2001)
4. M. Cerf, J. Harel, W. Einhäuser, C. Koch, Predicting human gaze using low-level saliency combined with face detection, in *Advances in Neural Information Processing Systems* (2008), pp. 241–248
5. I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato, A. Sugimoto, Head direction estimation from low resolution images with scene adaptation. Comput. Vis. Image Underst. **117**(10), 1502–1511 (2013)

6. D. Cornish, D. Dukette, *The Essential 20: Twenty Components of an Excellent Health Care Team* (Dorrance Publishing Co. Inc., 2010)
7. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in *CVPR (1)* (IEEE Computer Society, 2005), pp. 886–893
8. D. Das, M.M. Hoque, T. Onuki, Y. Kobayashi, Y. Kuno, Vision-based attention control system for socially interactive robots, in: *IEEE International Symposium on Robot and Human Interactive Communication* (Paris, France, 2012), pp. 496–502
9. D. Das, M.G. Rashed, Y. Kobayashi, Y. Kuno, Supporting human-robot interaction based on the level of visual focus of attention, in *IEEE Transactions on Human-Machine Systems* (Accepted for Publication)
10. G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avrithis, Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. IEEE Trans. Multimedia **15**(7), 1553–1568 (2013)
11. Y. Furukawa, J. Ponce, Accurate, dense, and robust multiview stereopsis. IEEE Trans. Pattern Anal. Mach. Intell. **32**(8), 1362–1376 (2010)
12. A.J. Glenstrup, T. Engell-Nielsen, Eye controlled media: present and future state. Ph.D. thesis, Information Psychology, University of Copenhagen, DIKU, DK-2100, Denmark, 1995
13. A. Hagiwara, A. Sugimoto, K. Kawamoto, Saliency-based image editing for guiding visual attention, in *Proceedings of the 1st International Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction* (ACM, 2011), pp. 43–48
14. A. Hagiwara, A. Sugimoto, K. Kawamoto, Saliency-based image editing for guiding visual attention, in *Proceedings of the 1st International Workshop on Pervasive Eye Tracking & #38; Mobile Eye-based Interaction, PETMEI '11* (ACM, New York, 2011), pp. 43–48
15. Y.S. Hajime Hata Hideki Koike, Visual attention guidance using image resolution control. J. Inf. Proc. Soc. Jpn. **56**(4), 1142–1151 (2015)
16. D.W. Hansen, Q. Ji, In the eye of the beholder: a survey of models for eyes and gaze. IEEE Trans. Pattern Anal. Mach. Intell. **32**(3), 478–500 (2010)
17. J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in *Advances in Neural Information Processing Systems* (2006), pp. 545–552
18. M.M. Hoque, D. Das, T. Onuki, Y. Kobayashi, Y. Kuno, An integrated approach of attention control of target human by nonverbal behaviors of robots in different viewing situations, in *IROS* (IEEE, 2012), pp. 1399–1406
19. M.M. Hoque, T. Onuki, Y. Kobayashi, Y. Kuno, Effect of robot's gaze behaviors for attracting and controlling human attention. Adv. Robot. **27**(11), 813–829 (2013)
20. L. Itti, P. Baldi, Bayesian surprise attracts human attention. Vis. Res. **49**(10), 1295–1306 (2009)
21. L. Itti, N. Dhavale, F. Pighin, Realistic avatar eye and head animation using a neurobiological model of visual attention, in *Optical Science and Technology, SPIE's 48th Annual Meeting* (International Society for Optics and Photonics, 2004), pp. 64–78
22. L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
23. T. Joachims, Making large-scale svm learning practical, in *Advances in Kernel Methods—Support Vector Learning* (1999)
24. J.C. Karremans, W. Stroebe, Beyond vicary's fantasies: the impact of subliminal priming and brand choice. J. Exp. Soc. Psychol. 792–798 (2006)
25. D. Kahneman, *Attention and Effort* (Prentice-Hall, 1973)
26. Y. Kim, A. Varshney, Persuading visual attention through geometry. IEEE Trans. Visual. Comput. Graph. **14**(4), 772–782 (2008)
27. H. Kobayashi, S. Kohshima, Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye. J. Hum. Evol. **40**, 419–435 (2001)
28. K. Liang, Y. Chahir, M. Molina, C. Tijus, F. Jouen, Appearance-based gaze tracking with spectral clustering and semi-supervised gaussian process regression, in *ETSA* (2013), pp. 17–23

29. F. Lu, T. Okabe, Y. Sugano, Y. Sato, Learning gaze biases with head motion for head pose-free gaze estimation. Image Vis. Comput. **32**(3), 169–179 (2014)
30. F. Lu, Y. Sugano, T. Okabe, Y. Sato, Adaptive linear regressionfor appearance-based gaze estimation. IEEE Trans. Pattern Anal. Mach. Intell. **10**, 2033–2046 (2014)
31. F. Martinez, A. Carbone, E. Pissaloux, Gaze estimation using local features and non-linear regression, in *ICIP* (2012), pp. 1961–1964
32. J. Nakajima, A. Kimura, A. Sugimoto, K. Kashino, Visual attention driven by auditory cues—selecting visual features in synchronization with attracting auditory events, in *MultiMedia Modeling—21st International Conference, MMM 2015* (Sydney, NSW, Australia, January 5–7, 2015), *Proceedings, Part II* (2015), pp. 74–86
33. J. Nakajima, A. Sugimoto, K. Kawamoto, Incorporating audio signals into constructing a visual saliency map, in *Image and Video Technology* (Springer, 2014), pp. 468–480
34. B. Noris, K. Benmachiche, A. Billard, Calibration-free eye gaze direction detection with gaussian processes, in *VISAPP* (2008), pp. 611–616
35. T. Onuki, K. Ida, T. Ezure, T. Ishinoda, K. Sano, Y. Kobayashi, Y. Kuno, Designing robot eyes and head and their motions for gaze communication. Int. Conf. Intell. Comput. (ICIC2014) **LNCS8588**, 607–618 (2014)
36. T. Onuki, T. Ishinoda, E. Tsuburaya, Y. Miyata, Y. Kobayashi, Y. Kuno, Designing robot eyes for communicating gaze. Interact. Stud. **14**(3), 451–479 (2013)
37. C.E. Rasmussen, C.K.I. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, 2006)
38. M. Rolf, M. Asada, Visual attention by audiovisual signal-level synchrony, in *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction Workshop on Attention Models in Robotics: Visual Systems for Better HRI* (2014)
39. J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, R. Pfeifer, Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub, in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on* (IEEE, 2008), pp. 962–967
40. B. Schauerte, B. Kühn, K. Kroschel, R. Stiefelhagen, Multimodal saliency-based attention for object-based scene analysis, in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on* (IEEE, 2011), pp. 1173–1179
41. B. Schauerte, R. Stiefelhagen, "wow!" bayesian surprise for salient acoustic event detection, in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (IEEE, 2013), pp. 6402–6406
42. R. Stiefelhagen, J. Yang, A. Waibel, Modeling focus of attention for meeting indexing based on multiple cues. IEEE Trans. Neural Netw. **13**(4), 928–938 (2002)
43. W. Stroebe, The subtle power of hidden messages. Sci. Am. Mind **23**, 46–51 (2012)
44. Y. Sugano, Y. Matsushita, Y. Sato, Appearance-based gaze estimation using visual saliency. IEEE Trans. Pattern Anal. Mach. Intell. **35**(2), 329–341 (2013)
45. Y. Sugano, Y. Matsushita, Y. Sato, Learning-by-synthesis for appearance-based 3d gaze estimation, in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)* (IEEE, 2014), pp. 1821–1828
46. K. Tan, D. Kriegman, N. Ahuja, Appearance-based eye gaze estimation, in *WACV* (2002), pp. 191–195
47. X. Tan, L. Qiao, W. Gao, J. Liu, Robust faces manifold modeling: most expressive versus most Sparse criterion, in *ICCV Workshops* (2010), pp. 139–146
48. F. Tarrés, Gtav face database, http://gps-tsc.upc.es/GTAV/ResearchAreas/UPCFaceDatabase/GTAVFaceDatabase.htm
49. A. Wagner, J. Wright, A. Ganesh, Z. Zhou, Y. Ma, Towards a practical face recognition system: robust registration and illumination by sparse representation. CVPR **2009**, 597–604 (2009)
50. C. Ware, *Information Visulization: Perception for Design* (Morgan Kaufmann Publishers Inc., San Francisco, 2004)
51. O. Williams, A. Blake, R. Cipolla, Sparse and semi-supervised visual mapping with the S^3GP, in *CVPR* (2006), pp. 230–237
52. J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation. PAMI **31**(2), 210–227 (2008)