


Article

# EmoTour: Estimating Emotion and Satisfaction of Users Based on Behavioral Cues and Audiovisual Data

Yuki Matsuda <sup>1,2,3\*</sup> , Dmitrii Fedotov <sup>4,5</sup>, Yuta Takahashi <sup>1</sup>, Yutaka Arakawa <sup>1,6</sup>, Keiichi Yasumoto <sup>1,3</sup> and Wolfgang Minker <sup>4</sup>

<sup>1</sup> Graduate School of Information Science, Nara Institute of Science and Technology, Nara 630-0192, Japan; takahashi.yuta.to2@is.naist.jp (Y.T.); ara@is.naist.jp (Y.A.); yasumoto@is.naist.jp (K.Y.)

<sup>2</sup> Fellow of Japan Society for the Promotion of Science, Tokyo 102-0083, Japan

<sup>3</sup> RIKEN, Center for Advanced Intelligence Project AIP, Tokyo 103-0027, Japan

<sup>4</sup> Institute of Communications Engineering, Ulm University, 89081 Ulm, Germany; dmitrii.fedotov@uni-ulm.de (D.F.); wolfgang.minker@uni-ulm.de (W.M.)

<sup>5</sup> ITMO University, Saint Petersburg 197101, Russia

<sup>6</sup> JST Presto, Tokyo 102-0076, Japan

\* Correspondence: yukimat.jp@gmail.com; Tel.: +81-743-72-5392

Received: 16 October 2018; Accepted: 12 November 2018; Published: 15 November 2018

**Abstract:** With the spread of smart devices, people may obtain a variety of information on their surrounding environment thanks to sensing technologies. To design more context-aware systems, psychological user context (e.g., emotional status) is a substantial factor for providing useful information in an appropriate timing. As a typical use case that has a high demand for context awareness but is not tackled widely yet, we focus on the tourism domain. In this study, we aim to estimate the emotional status and satisfaction level of tourists during sightseeing by using unconscious and natural tourist actions. As tourist actions, behavioral cues (eye and head/body movement) and audiovisual data (facial/vocal expressions) were collected during sightseeing using an eye-gaze tracker, physical-activity sensors, and a smartphone. Then, we derived high-level features, e.g., head tilt and footsteps, from behavioral cues. We also used existing databases of emotionally rich interactions to train emotion-recognition models and apply them in a cross-corpus fashion to generate emotional-state prediction for the audiovisual data. Finally, the features from several modalities are fused to estimate the emotion of tourists during sightseeing. To evaluate our system, we conducted experiments with 22 tourists in two different touristic areas located in Germany and Japan. As a result, we confirmed the feasibility of estimating both the emotional status and satisfaction level of tourists. In addition, we found that effective features used for emotion and satisfaction estimation are different among tourists with different cultural backgrounds.

**Keywords:** ubiquitous computing; emotion recognition; satisfaction estimation; wearable computing; dialogue systems; smart tourism; smart cities

## 1. Introduction

Due to the ubiquity of smart devices, including smartphones and wearables, people can find various and helpful pieces of real-time living environment information, such as about the weather and roadway traffic. Moreover, some of the recently emerged systems account for user context, e.g., recording what/where users are doing at that moment. To design more context-aware systems, psychological user context (e.g., emotional status) needs to be taken into account since it differs across users, and even for the same user at different times. In this study, we focus on the tourist domain. It is a typical use case with high demand for context awareness. In fact, the emotional status and

satisfaction level of tourists are susceptible during sightseeing; hence, observing emotional feedbacks is useful for providing context-aware tourist guidance.

The aim of our study is to estimate the emotional status and satisfaction level of users susceptible to rating their activities (e.g., tourists during sightseeing). Recently, various approaches have been proposed to understand users' emotional status and satisfaction level in order to use them for consumer services. The most widely used approaches to collect satisfaction level of users are online user reviews and questionnaires [1,2]. User reviews are also used in consumer services, such as the rating systems of TripAdvisor [3], Yelp [4], and Amazon [5]. However, keeping users motivated to regularly write reviews is difficult, especially with medium rating values, which leads to the risk of biased distribution of review ratings. To provide reliable information for other users, it is necessary to collect quantitative data without user reviews. Though many studies have tried to estimate the emotional status of users with methods based on audiovisual-data analysis [6–12], the accuracy of emotion recognition in outdoor places, such as tourist sights, tends to be worse due to the inclusion of environmental noise in audiovisual data [13–15]. In recent studies, the range of modalities has been expanding to physiological features (e.g., body movement, eye gaze) [16–24], and a fusion of them might help with even outdoor estimation [17,25,26].

In this article, we propose a tourist emotion- and satisfaction-estimation system named EmoTour. EmoTour employs several kinds of modalities taken from actions that tourists naturally and unconsciously do during sightseeing. Since tourists often naturally record videos or take pictures by themselves (e.g., a selfie), and also unconsciously behave (e.g., walk around, gaze at an object) during sightseeing, audiovisual data and behavioral cues are used as modalities. In our previous paper, we have already confirmed that behavioral cues have a relationship with the emotion and satisfaction of tourists [27]. With EmoTour, we have built a model for estimating tourists' emotion and satisfaction during sightseeing by fusing features derived from each modality.

The main contributions of this paper are the following:

- First, we propose a new model for quantitatively estimating both the emotion and satisfaction of tourists by employing multiple modalities obtained from unconscious and natural user actions. To avoid the potential risk of biased ratings in a user review for satisfaction-level estimation, and enable emotional-state estimation at an actual sightseeing situation, we employ the combination of behavioral cues and audiovisual data collected by an eye-gaze tracker, physical-activity sensors, and a smartphone. In detail, the following high-level features were derived from each modality and fused to build a final classifier: eye movement, head tilt, and footsteps from behavioral cues; and vocal and facial expressions from audiovisual data. We argue that our scheme can build the model without dependence on any extra tasks for users.
- Second, we evaluated our model through experiments with 22 users in a tourist domain (i.e., in a real-world scenario). As the experimental fields, we selected two touristic areas, located in Germany and Japan, which have completely different conditions. We evaluated the emotion estimation model through a three-class classification task (positive, neutral, negative) using unweighted average recall (UAR) score as a metric, and achieved up to 0.48 of UAR score. Then, we evaluated the satisfaction estimation model through a 7-level regression task (0: fully unsatisfied–6: fully satisfied) using mean absolute error (MAE) as a metric, and achieved up to 1.11 of MAE. In addition, we found that effective features used for emotion and satisfaction estimation are different among tourists with different cultural background.

The rest of the paper is organized as follows. Section 2 introduces the current status of related studies and services, and defines the challenges that should be overcome in our study. Section 3 considers the approach for estimating user emotion and satisfaction, and explains the concrete workflow of our approach and the modalities used. Section 4 describes the methodology of user emotion and satisfaction estimation including feature extraction and modality fusion. Moreover, we evaluate our method through real-world experiments in Section 5, and discuss the contribution

and limitation of our method in Section 6. Finally, conclusions and suggestions for future work are given in Section 7.

## 2. Related Work and Challenges

Due to the high demands of context-aware systems, especially in the tourist domain, there are many studies focused on environmental sensing to collect real-time information of tourist sights [28,29]. On the other hand, the estimation of the psychological user context, which may be used for providing appropriate information based on the situation, has not yet been deeply tackled in spite of its importance.

Our motivation was the enhancement of a context-aware tourist-guidance system by introducing psychological context estimation in addition to existing environmental sensing technologies. In the following sections, we describe related works that widely include other domains, then clarify the objective and challenges of our study, and introduce our preliminary work.

### 2.1. Estimation of Emotional Status

Emotion recognition has been a hot topic for many research areas and for several years now due to the high demand for context-aware systems, such as spoken-dialogue systems. However, there are not many studies targeting the tourist domain yet.

In emotion recognition, audio- and/or visual-based approaches are popular fields in the field of dialogue systems and human-computer interaction [30]. In laboratory (indoor) conditions, existing audio-based emotion-recognition systems that use a deep neural network have achieved great performance [6,7]. Quirk et al. proposed an audio-based emotion-recognition system [8]. They built a dialogue system on mobile devices, and achieved around 60% recall score for four affective dimensions. Tarnowski et al. proposed an approach based on facial movements [9]. They obtained good classification accuracy of 73% for seven facial expressions. Moreover, they mentioned that head movements (orientation) could significantly affect extracting facial-expression features. Aiming at higher accuracies, bimodal emotion-recognition methods, combining audio and visual features, were also proposed [10,11], and they achieved better accuracy (e.g., 91% for six emotion classes). However, in outdoor conditions, the accuracy of emotion recognition tends to be worse due to the inclusion of environmental noise in audiovisual data. According to Emotion Recognition in the Wild Challenge 2017 (EmotiW), the accuracy of emotion recognition is up to 60% for seven emotion classes [13–15]. Since tourist sights may have noisier environments, we should take such environmental conditions into account to estimate the emotional status of tourists.

To infer the emotion of a person, the unconscious behavior of humans may be a clue as well. Shapsough et al. described that emotions could be recognised by using typing behaviour on the smartphone [16]. This approach used a machine-learning technique and induced high accuracy on emotion recognition, yet it is not feasible to frequently ask users to type on their smartphone during a sightseeing tour. Resch et al. proposed an emotion-collecting system for urban planning called Urban Emotions [17]. The paper describes that wrist-type wearable devices and social media were used for emotion measurements. Since this approach relies on an assumption that posts on social media are written in situ, it has the problem of spatial coverage for collecting data. In recent large social media services (e.g., Twitter, Facebook), users cannot attach exact location data to a post with default settings, which makes UrbanEmotions difficult to collect comprehensive data. However, it also suggested that body movement can be used for recognising emotion.

Moreover, recent studies in the field of emotion recognition focus on expanding the range of modalities and combining them. Ringeval et al. proposed to introduce physiological features in addition to audiovisual ones, and to build a multimodal system that relies on their combination [12]. As physiological features, an electrocardiogram and electrodermal activity were used. Physiological features provided lower performance and weaker correlation than audiovisual ones with continuous emotional labels, but helped to increase the overall performance of a multimodal system. Many studies also introduce physiological features: heart-related (electrocardiogram, heartbeat), skin-

and blood-related (electrodermal activity, blood-pressure), brain-related (electroencephalography), eye-related (eye gaze, pupil size), and movement-related (gestures, gyrosopic data) [17–24], and, in many cases, improvement of accuracy was observed, even in outdoor conditions [17,25,26].

## 2.2. Estimation of Satisfaction Level

In current consumer services, such as TripAdvisor [3], Yelp [4], and Amazon [5], online user reviews and questionnaires are still widely used to collect the satisfaction level of users. TripAdvisor [3] in particular uses a five-star rating system and comments from tourists as user reviews about sightseeing spots. To guarantee quantity and quality of voluntary reviews, it is essential to provide a motivation to contributors. However, keeping users motivated to regularly write reviews is difficult, especially with medium rating values, because many people do not like to post a review without any external incentives when they felt “there’s nothing special”. It means there is a risk of skewing the evaluation due to an imbalance of reviewers’ distribution.

Many studies also adopt questionnaire-based surveys for measuring tourist satisfaction [1,2]. Fundamentally, several hundred samples (respondents) are required to produce reliable data. However, because the questionnaire-based method relies on the manual tasks of a human, it has problems in sustainability and the spatial coverage of the survey. It also has the same risk as the method based on user reviews and ratings.

## 2.3. Objective and Challenges

Our objective was to determine the quantitative emotional status and satisfaction levels of users to design more intelligent and reliable guidance systems. However, through the investigation of current studies and services, we found several problems (e.g., biased reviews, spatial coverage of evaluation, accuracy of estimation) for applying existing techniques to real conditions of use.

From this background, the main challenge of our study was to establish a state-of-the-art method for estimating user emotion and satisfaction by fusing audiovisual data and various sensor data specialising in user behavior. In the following sections, we describe the design and implementation of a method for estimating the emotional status and satisfaction level of tourists, and provide deeper evaluation and discussion through real-world experiments.

## 2.4. Preliminary Work

As our preliminary work, we found a correlation between user psychological context (emotional status, satisfaction level) and user unconscious behavior [27]. Then, we proposed a basic setup for estimating the emotional user status, and confirmed the feasibility of the concept of our study through a small experiment in real-world conditions [31]. In this paper, we explain the method for estimating not only the emotional status but also the satisfaction level in detail, and evaluate both methods through experiments with additional participants (nine people added, 1.7 times the scale). Moreover, we analyzed the effects of differences in cultural backgrounds of tourists on the accuracy of estimation, and discuss future perspectives.

## 3. Proposed Approach and Workflow

Our approach was designed for the tourist domain, where users are walking on a path through different sightseeing areas. To estimate the emotion and satisfaction of tourists during sightseeing, we focus on various actions that tourists unconsciously and naturally do during sightseeing. For example, tourists might approach a scenic place or work of art, stop, and gaze at it, potentially take selfie photos, or send video messages to their friends. The accumulation of these natural actions should be linked to the emotional status and satisfaction level that they feel there. Through a preliminary study, we have already confirmed that several actions (eye gaze and head/body movement) have a relationship with the emotion and satisfaction of tourists [27]. Hence, we propose an approach

to observe such natural actions of tourists and estimate the tourist emotion/satisfaction while performing them.

Figure 1 shows the workflow of our whole system for collecting the data of tourist actions and labels. The overview of each step is as follows:

### Step 1—Split the whole tour into sessions

Before starting sightseeing, we split the whole tour into small periods (sessions) that included at least one sight each. We assumed that a tourist typically requests guidance information for each sightseeing spot.

### Step 2—Sensing and labeling

Tourists could freely visit sights while equipped with wearable devices that continuously recorded their behavior during the whole sightseeing. At the end of each session, they gave small amounts of feedback about the latest session by recording a selfie video. We assumed that recording a video serves as a means of interacting with dialogue systems or sending a video message to their friends. They also manually input their current emotional status and satisfaction level as a label. Then, they repeated the same procedure for each of the tour sessions.

### Step 3—Building the estimating model

The tourist emotion- and satisfaction-estimation model was built based on tourist behavior, audiovisual data, and labels.

In the following sections, we describe the details of the modalities and labels.

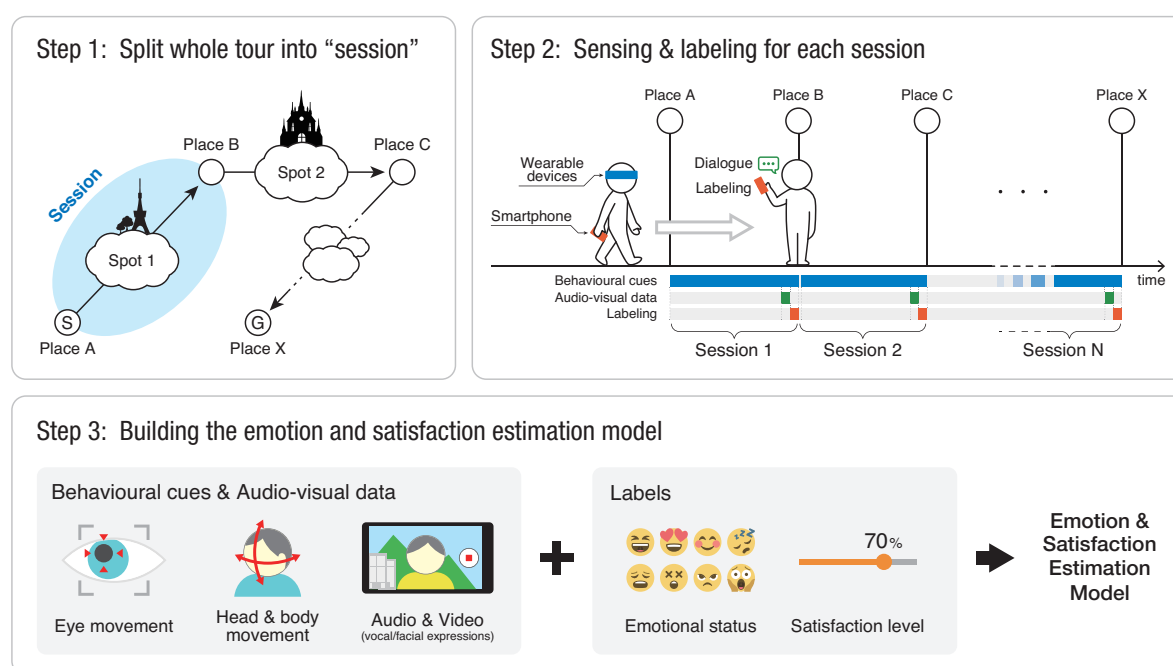


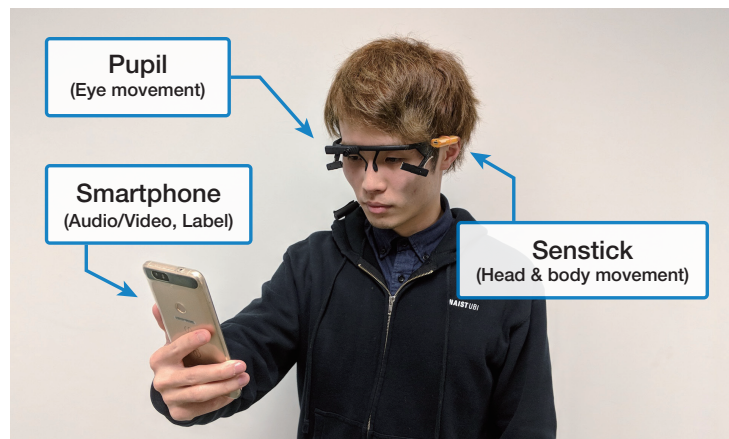
Figure 1. Workflow to estimate tourist emotion and satisfaction level.

#### 3.1. Modalities

To perform an emotion estimation in the tourist domain, we used multimodal features: audiovisual data (vocal/facial expressions) and behavioral cues (eye and head/body movement data). Since tourists often take videos or photos, e.g., a selfie, audiovisual data could be used for our study. However, accuracy may have been low due to environmental issues in outdoor places, as mentioned in Section 2. Hence, we additionally used the features extracted from various tourist behaviors that happen unconsciously during sightseeing. The sense of sight is one of the most important sensory systems in sightseeing, and it can be tracked as eye movement using existing wearable devices and technologies. Moreover, due to the directivity on sensory systems (e.g., hearing, sight), head and body

movements may be affected by them. Thus, we used head and body movements as features in addition to eye movement.

In our study, we used the three devices shown in Figure 2 to record features in real time: an Android smartphone (GPS-data, audiovisual data), a mobile eye-tracking headset Pupil with two 120 Hz eye cameras [32] (eye gaze, pupil features), and a sensor board SenStick [33] mounted on an ear of the eye-tracking device (accelerometer, gyroscope).



**Figure 2.** Devices for data collection during sightseeing: pupil [32], SenStick [33], and smartphone.

### 3.2. Labels

To represent the psychological context of tourists, we employed two types of metrics: emotional status and satisfaction level. We collected these data as labels by using the Android application shown in Figure 3. Tourists could manually enter the ratings of the session at the end of each session. The details of each metric are described as follows:

#### Emotional status

To represent the emotional status of tourists, we adopted the two-dimensional map defined on Russell's circumplex space model [34]. Figure 4 shows the representation of the emotional status. We divided this map into nine emotion categories and classified them into three emotion groups as follows:

**Positive** : Excited (0), Happy/Pleased (1), Calm/Relaxed (2)

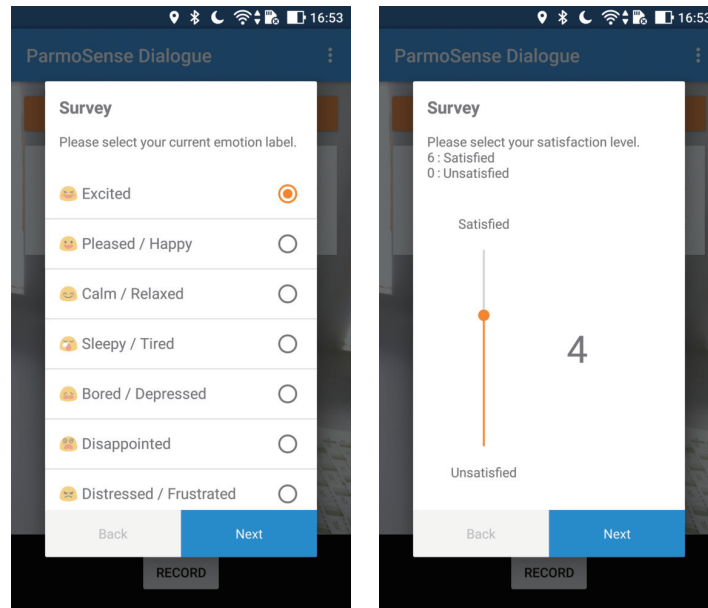
**Neutral** : Neutral (3)

**Negative** : Sleepy/Tired (4), Bored/Depressed (5), Disappointed (6),  
Distressed/Frustrated (7), Afraid/Alarmed (8)

As a side note, Russell's model [34] is mainly employed for the time-continuous annotation of audiovisual databases that we used to build pre-trained models in Section 4.

#### Satisfaction level

To represent the satisfaction level of tourists, we used the Seven-Point Likert scale which the Japanese government (Ministry of Land, Infrastructure, Transport, and Tourism) uses as the official method. Tourists could choose their current satisfaction level between 0 (fully unsatisfied) and 6 (fully satisfied). A neutral satisfaction level is 3 and it should approximately represent the state of the participant at the beginning of the experiment.



(a) Emotional state. (b) Satisfaction level.

Figure 3. Smartphone application for collecting labels from tourists.

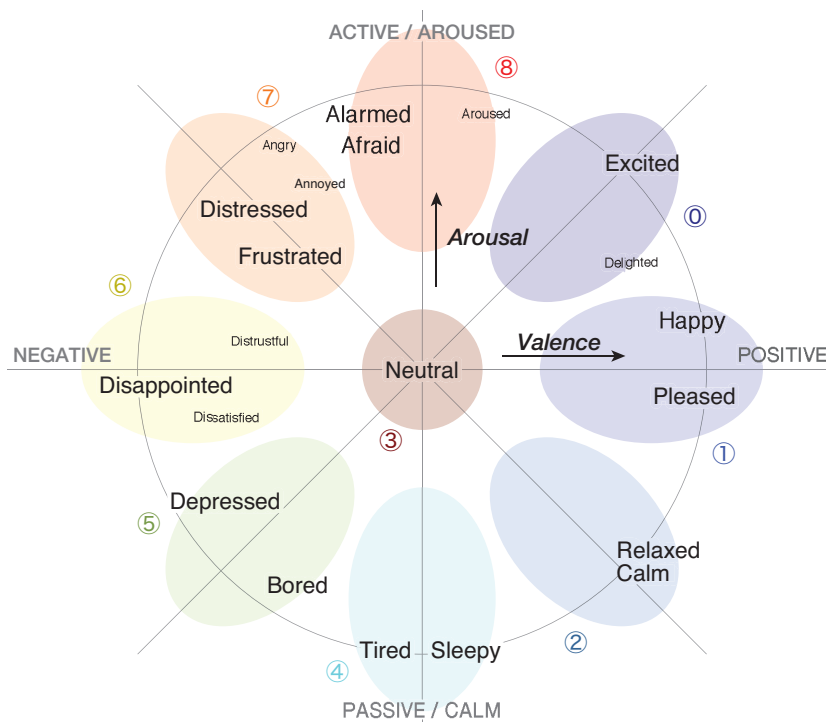
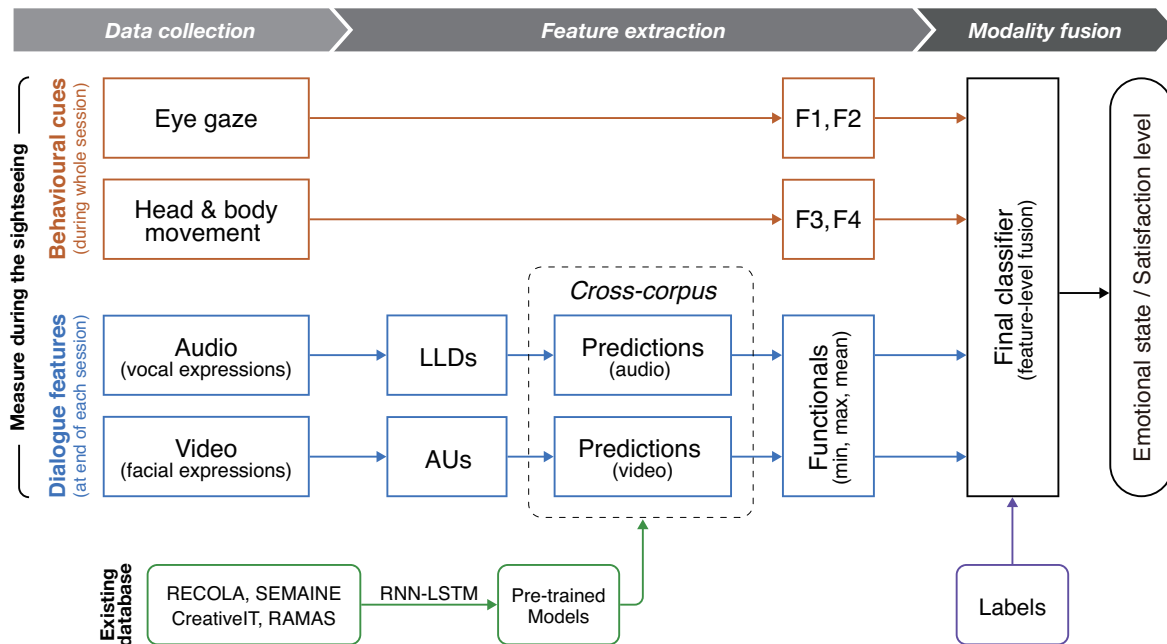


Figure 4. A two-dimensional emotion status model. Figure is taken from References [34,35].

#### 4. Methodology of Tourist Emotion and Satisfaction Estimation

In this section, we describe the methodology of estimating emotion and satisfaction of tourists. Figure 5 depicts the scheme of our method, which consists of three stages. The first stage is Data Collection described in Section 3. The second stage is Feature Extraction, described in Section 4.1, where we preprocessed the collected raw data and extracted several features for each modality. The final stage is Modality Fusion described in Section 4.2, where several modalities were combined to build the final classifier of tourist emotion and satisfaction estimation.



**Figure 5.** Scheme of tourist emotion and satisfaction estimation with modality fusion. LLDs: low-level descriptors of prosodic, spectral, cepstral, and voice quality, AUs: action units for describing facial expressions, F1: intensity of eye movement, F2: statistical features of eye movement, F3: head movement (head tilt), F4: body movement (footsteps), RNN-LSTM: recurrent neural network with long short-term memory.

#### 4.1. Preprocessing and Feature Extraction

The raw data from each modality cannot be directly used to build tourist emotion and satisfaction estimation model. Hence, the methods of data preprocessing and feature extraction explained in Section 4.1.1 (behavioural cues) and Section 4.1.2 (audio-visual data) are applied to each modality.

##### 4.1.1. Behavioral Cues—Eye-, Head-, and Body-Movement Features

Eye-movement features were extracted using the Pupil Labs eye tracker [32]. We used *theta* and *phi* values, which represent a normal pupil as a 3D circle in spherical co-ordinates (Figure 6a). Hence, we could only use two variables to describe the position of pupils and, thus, the eye gaze. Note that the raw values of eye-movement data differ across users and depend on the physical setting of camera and eye peculiarity. The eye-gaze data were analyzed using the following methodologies:

##### F1: Intensity of eye movement

Minimum and maximum values for *theta* and *phi* were calculated for each participant; eight thresholds (10%–90%, 10% step, except 50%) were set for the range [min, max] as shown in Figure 6b, and then used to count the percentage of time outside each threshold per session. In total, 16 features were used.

##### F2: Statistical features of eye movement

Average and standard deviation of *theta* and *phi* were calculated for a small window of recorded data and the values corresponding to the same session were averaged. The following window sizes were used: 1, 5, 10, 20, 60, 120, 180, and 240 s with the offset of  $\frac{1}{3}$  of the window size. In total, 64 features were used.



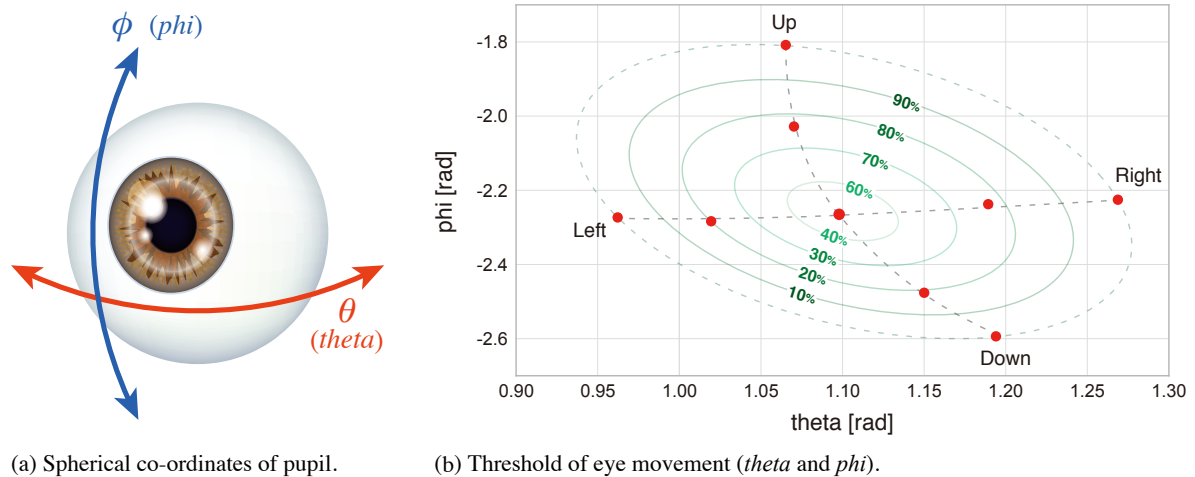


Figure 6. Representation of eye movement.

Then, head and body movement features were extracted using the inertial-sensor (accelerometer and gyroscope) values of the SenStick [33]. Both sensors have three axes: the X-axis, Y-axis, and Z-axis. Head- and body-movement data were analyzed using the following methodologies:

#### F3: Head movement (head tilt)

As a head movement, head tilt was derived using gyroscope values. The average  $\mu$  and the standard deviation  $\sigma$  of the gyroscope values were calculated for each participant. Then, the upper/lower thresholds  $\psi$  were set with the following equations (Equations (1) and (2)). The parameter  $a$  represents the axis of the gyroscope.

$$\psi_{upper,a} = \mu_a + 2\sigma_a, \quad (1)$$

$$\psi_{lower,a} = \mu_a - 2\sigma_a. \quad (2)$$

Finally, head tilt (looking up/down, right/left) was detected using threshold  $\psi$ . In our condition, the Y-axis indicates a looking-up/down motion, and the Z-axis indicates a looking-left/right motion. Since the duration of each session was different, we converted these data to several features: head tilt per second; and average and standard deviation of the time interval looking at each direction. In total, 23 features were used.

#### F4: Body movement (footsteps)

Footsteps are analyzed with a method based on the approach of Ying et al. [36]. First, the noises of accelerometer values were removed by applying a Butterworth filter with 5 Hz cutoff frequency. Then, high-frequency components were emphasised through the differential processing shown in Equation (3). The parameter  $x(n)$  represents the accelerometer value at index  $n$ .

$$y(n) = \frac{1}{8} \{2x(n) + x(n-1) - x(n-3) - 2x(n-4)\}. \quad (3)$$

Furthermore, the following integration process (Equation (4)) smoothed the accelerometer values, and small peaks of them were removed. In our condition,  $N$  was chosen to be 5 empirically. Since the sensor position was different from the original method in our condition, we used a modified parameter.

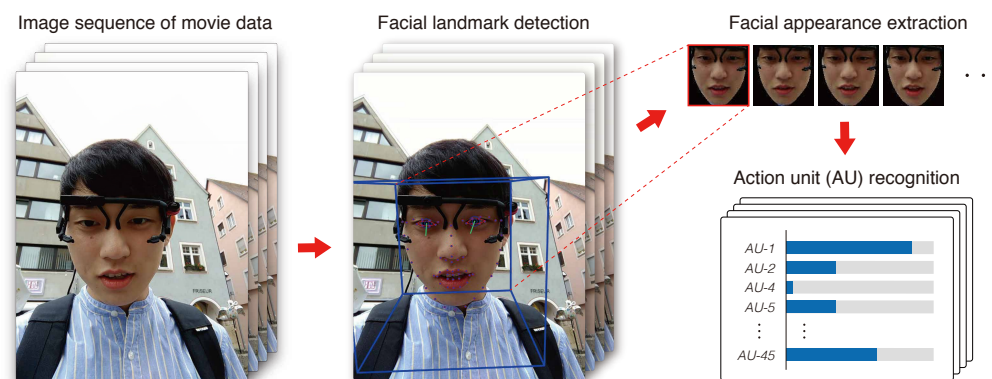
$$y(n) = \frac{1}{N} \{x(n-(N-1)) + x(n-(N-2)) + \dots + x(n)\}. \quad (4)$$

Finally, footsteps were extracted by counting local maximum points. As features, we used footsteps per second, and average and standard deviation of a time interval for each step. In total, five features were used.

#### 4.1.2. Audiovisual Data—Vocal and Facial Expressions

Audio features (vocal expressions) were extracted with openSMILE software [37]. They consisted of 65 low-level descriptors (LLDs) of four different groups (prosodic, spectral, cepstral, and voice quality) and their first-order derivatives (130 features in total) used in ComParE challenges since 2013 [38]. Window size was set to 60 ms, and window step size was set to 10 ms, resulting in feature extraction on overlapping windows with a rate of 100 Hz.

As video features (facial expressions), action units (AUs) were extracted with OpenFace [39,40], an open-source toolkit. AUs describe specific movements of facial muscles in accordance with the facial action coding system (FACS) [41,42], e.g., AU-1 means an action of “raising up the inner brow.” We used the following 17 AUs, which can be extracted using OpenFace: 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, 45. The procedure of AU extraction is shown in Figure 7. First, the movie file is converted to a sequence of images, and facial landmarks are detected for each frame. Then, the face images are clipped out from the original frames because it expects that background objects and random people are partially captured in movie data in outdoor places. Finally, AUs are recognized by fusing several features taken from a clipped image for each frame of an original movie. In total, 18 AU-related features were extracted.



**Figure 7.** Visual features (action units) extraction using Openface [39,40].

Due to the lack of data for training a decent audiovisual-based emotion-recognition system, in this study we used models that are trained in advance by utilizing several existing corpora of emotionally rich interactions: the RECOLA (Remote COLlaborative and Affective interactions) database [12], SEMAINE (Sustained Emotionally coloured MACHine-human Interaction using Nonverbal Expression) database [43], CreativeIT database [44], and RAMAS (The Russian Acted Multimodal Affective Set) database [45]. As corpora have different annotation rates, they were brought to the same data frequency to be able to share the same prediction models. The least frequency of 25 Hz, presented in RECOLA, was used for the remaining corpora.

The models are based on recurrent neural networks with long short-term memory (RNN-LSTM) and built in a way to consider the particular amount of context (7.6 s) in accordance to our previous study [46], as it shows better results. Networks were comprised of two hidden LSTM layers of 80 and 60 neurons with ReLU (Rectified Linear Unit) activation function, respectively, each followed by a dropout layer with a probability of 0.3. The last layer has one neuron with linear activation function for regression tasks (databases: RECOLA, SEMAINE, CreativeIt) and six neurons with softmax activation function for the classification task (database RAMAS). We used RMSProp [47] as an optimizer with a learning rate of 0.01. For regression tasks, we utilized a loss function based on the concordance correlation coefficient that takes into account not only the correlation between two sets, but also the divergence, being not immune to biases. For the classification task, we used cross-categorical entropy.

After feeding the features extracted from the audiovisual data of our experiment to the trained model, for each time step we obtained the prediction in arousal and valence if we used the regression models, and probabilities of particular emotion if we used the classification models.

Models that are trained on the additional corpora cannot be directly used for emotional-status estimation in the context of our method due to the following reasons:

- Labels differ from those collected through our system in range and dimensions, i.e., they are on the arousal–valence scale instead of emotions for regression tasks, and an emotion set for a classification task does not match with ours.
- They are time-continuous, i.e., each value represents the emotional state for one frame of the audiovisual data, though we had one label per each session.

Taking these differences into account, predictions should be generalised and adapted to our method without losing valuable information. To achieve it, we took simple functionals (*min*, *max*, *mean*) from dimensional labels, i.e., arousal and valence as well as mean prediction scores for categorical emotions separately for each session, and merged them into feature vectors. Thus, we had high-level predictions from earlier trained models as features in our system.

For each modality, we used a simple feed-forward neural network with one hidden layer to make unimodal prediction from high-level features.

#### 4.2. Modality Fusion

To build our final tourist emotion- and satisfaction-estimation system, we combined predictions based on our features, described in Section 4.1, on two levels: feature and decision. During the experiment, some problems with the devices and the data-collection process occurred that led to some data missing. The final classifier should be robust and able to work with incomplete feature sets. On the decision-level fusion, it was achieved by applying linear models, where the final label is assigned, based on a linear combination of existing lower-level predictions. On the feature level, such feature sets were filled with zeros.

## 5. Experiments and Evaluation

### 5.1. Overview of Real-World Experiments

We conducted experiments in real-world conditions to evaluate the tourist emotion- and satisfaction-estimation method. As the experimental fields, we selected two tourist areas depicted in Figure 8 that have completely different conditions. The first one is the centre of Ulm, Germany. The sights in this area include particular buildings as well as walking routes with high tourist value (e.g., the Fisherman’s Quarter). The sights are surrounded by common city buildings and may be crowded depending on the time. The approximate length of the route is 1.5 km, divided into eight sessions. The second area is Nara Park, the historic outskirts of Nara, Japan. The route through the area includes many scenic and religious buildings (temples and shrines) that are located in nature, and has no distraction from the sights included in the sessions. The approximate length of the route is 2 km, divided into seven sessions.

Participants were asked to follow prepared routes and take as much time as they needed to see the sights. During the sessions, we recorded the data in real time according to Figure 1. At the end of each session, participants were asked to provide small feedback and labels for emotion and satisfaction level.

In total, we conducted our experiment with 22 participants, and collected 183 sessions’ data. The distribution of participants was the following: age range—22–31 years old (average age is 24.3); nationalities—12 Japanese, 10 Russian; gender—17 males, 5 females. In addition, 17 and 5 people went sightseeing to the tourist area located in Germany and Japan, respectively. Most of them were real tourists (e.g., short-term international students or new students, visitors), and hence they were not familiar with the experimental fields.

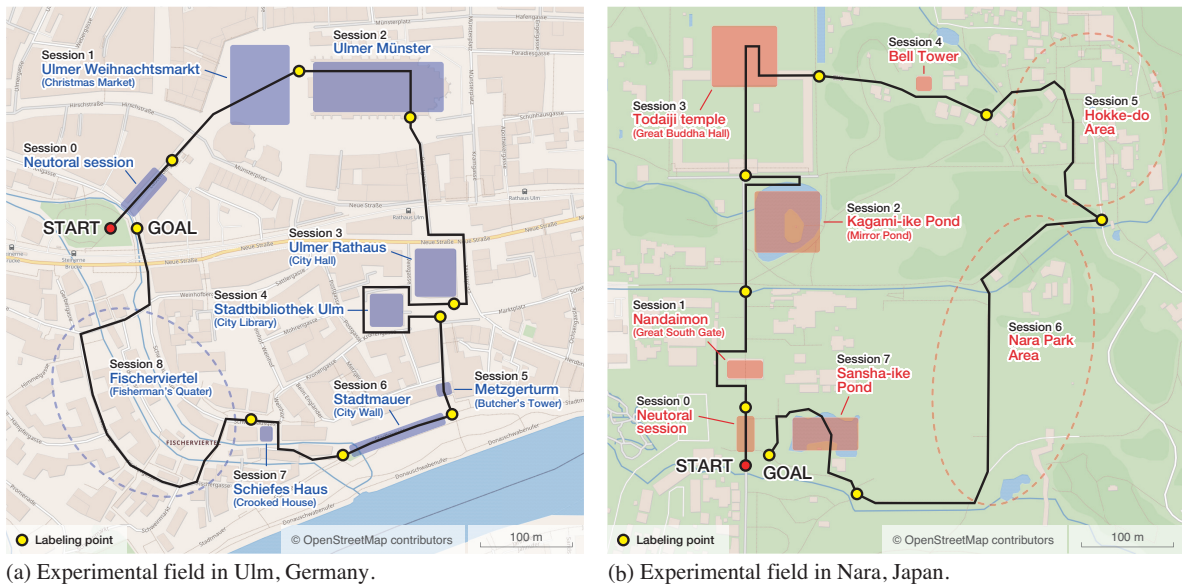


Figure 8. Experimental fields.

To collect the labels, the emotional status for each session during sightseeing could be measured only by the participants themselves. Due to the natural impression that sightseeing is something interesting, labels collected from tourists tend to be imbalanced. In our experiments, the distribution of labels was as shown in Figure 9a, and the ratio of each emotion group was: positive: 71.0%, neutral: 17.5%, negative: 11.5%. To manage this condition, we used UAR as the performance metric. The same imbalance is presented for the satisfaction level as shown in Figure 9b. Then, the relationship between emotion categories (groups) and satisfaction level is shown in Figure 10a,b, and they suggest a moderate positive correlation ( $r = 0.644$ ), which proves the agreement between labels.

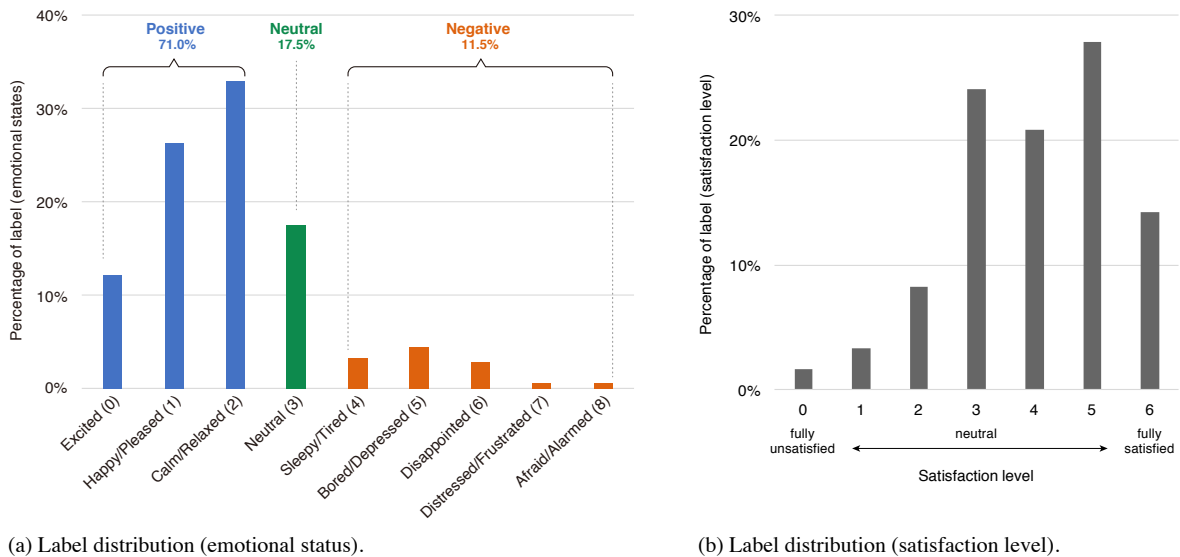
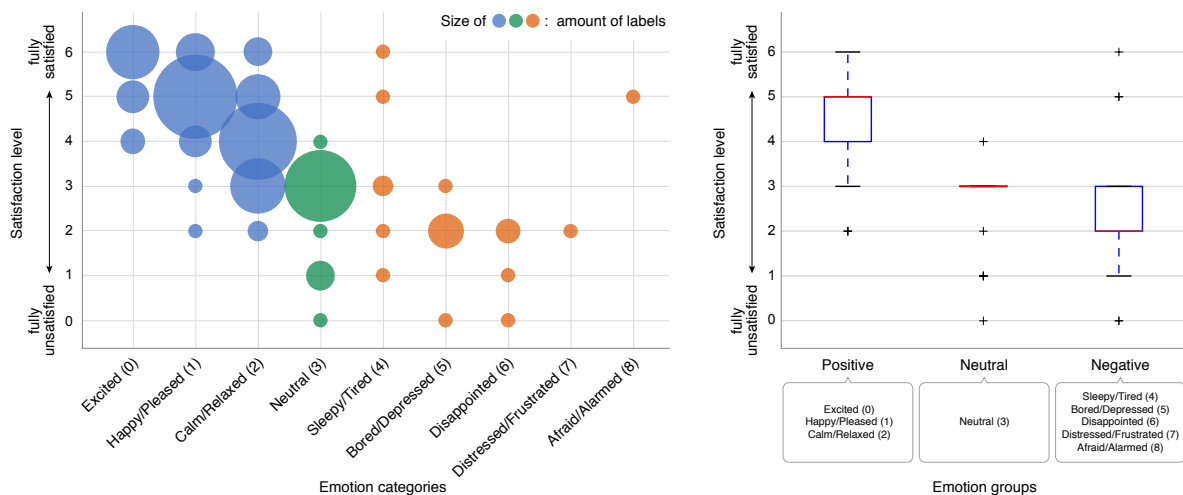


Figure 9. Label distribution.



(a) The distribution of labels (nine emotion categories).

(b) The distribution of labels (three emotion groups).

**Figure 10.** Relationship between emotional status and satisfaction level.

## 5.2. Results

To evaluate our emotion- and satisfaction-estimation system, we conducted a series of reproducible experiments. To provide a fair comparison of different modalities or fusion methods, we fixed random seeds, resulting in consistent sample shuffling. The results for uni- and multimodal emotion and satisfaction estimation are presented in Table 1. We applied two different fusion approaches: feature-level fusion, where we built one model on merged feature vectors from corresponding modalities; and decision-level fusion, where we used a linear combination of prediction scores from unimodal models to set the final label.

We evaluated the performance of our emotion estimation system through a classification task of three emotion groups: Positive (0–2), Neutral (3), and Negative (4–8). Due to imbalanced label distribution, we used UAR as the performance metric. This ranged from 0 to 1 (the higher the better) and, for three-class classification problems, it had a chance level of 0.33. The results (Table 1) show that our system performed at up to 0.48 of a UAR score in three-class (all emotions) emotion-estimation tasks. At the unimodal level, we found a performance of 0.45 using head/body-movement features (F3, F4). This suggests that there are implicit connections between emotional status and features, calculated for these modalities. Then, the highest performance of 0.48 was shown with all features (decision-level fusion), and it proves that combining behavioral cues and audiovisual data is useful to estimate the tourists' emotions.

Then, to evaluate the performance of satisfaction estimation, we derived the MAE between estimated value and label. The MAE score can be any positive value, and 0 represents a perfect match. Table 1 shows that our system performed at a low MAE up to 1.12 in seven-level satisfaction-estimation tasks. The highest performance was shown with all features (decision-level fusion) as with emotion-estimation tasks.

**Table 1.** Performance of uni- and multimodal tourist emotion and satisfaction estimation. The best performances are highlighted with bold text.

Modality	Emotion (Unweighted Average Recall: UAR)		Satisfaction (Mean Absolute Error: MAE)	
Eye movement (F1, F2)	0.401		1.238	
Head/body movement (F3, F4)	0.434		1.230	
<b>Behavioral cues (eye + head/body movement)</b>	0.458		1.265	
Audio (vocal expressions)	0.386		1.208	
Video (facial expressions)	0.411		1.198	
<b>Audiovisual data (audio + video)</b>	0.414		1.194	
<b>Feature-level fusion</b>	0.428		1.311	
<b>Decision-level fusion</b>	<b>0.484</b>		<b>1.110</b>	

In our experiment, participants could be roughly divided into two almost equal groups by their nationality: Japanese (12 people) and Russian (10 people). It was proven by previous studies that different culture groups express emotions differently in ways and intensity [48–50]. To see whether the difference in cultural background affected the accuracy of estimation in our study as well, we conducted the same modeling procedures, described above, with two culturally differentiated groups of participants. Then, to evaluate the performance of emotion and satisfaction estimation, the UAR score and MAE were derived, respectively.

The results are shown in Table 2, and they suggest that the impact of cultural difference exists, especially on emotion-estimation tasks. For Japanese tourists, the result shows that audiovisual data showed a better UAR score of 0.45 than behavioral cues (0.42) at the unimodal level. Then, we confirmed that their decision-level fusion showed the highest UAR score of 0.47. Interestingly, we found that there was an opposite characteristic for Russian tourists. The highest UAR score of 0.57 was shown by behavioral cues; in contrast, the low score of 0.34–0.41 was shown by audiovisual data, and video-based models performed almost at a chance level.

**Table 2.** Performance of tourist emotion and satisfaction estimation (by nationality of participants). The best performances are highlighted with bold text.

Modality	Emotion (UAR)		Satisfaction (MAE)	
	Japanese	Russian	Japanese	Russian
Eye movement (F1, F2)	0.438	0.426	1.045	1.345
Head/body movement (F3, F4)	0.417	0.438	1.314	1.290
<b>Behavioral cues (eye + head/body movement)</b>	0.415	<b>0.576</b>	1.099	1.347
Audio (vocal expressions)	0.447	0.372	1.093	1.304
Video (facial expressions)	0.463	0.346	1.100	1.300
<b>Audiovisual data (audio + video)</b>	0.445	0.417	1.067	1.300
<b>Feature-level fusion</b>	0.423	0.507	1.190	1.420
<b>Decision-level fusion</b>	<b>0.473</b>	0.496	<b>1.000</b>	<b>1.157</b>

These results suggest that we need to take the effects of nationalities or cultural differences into account in order to generalize the model of our system. As future work, we will expand the nationalities of tourists, and build a general model using them in consideration of their cultural background. In addition, we aim to investigate the effects of other attributes of tourists, such as gender and age.

## 6. Discussion and Limitations

### 6.1. Feasibility of Our Proposed System

The results from Section 5.2 prove that tourists' emotion and satisfaction estimation can be extracted by our proposed system to a certain degree. In our study, modality fusion showed a better result compared to unimodal systems for estimating both emotional status and satisfaction level. Especially combining behavioral features at a feature-level often improved the performance of emotion estimation, compared to eye- and head-based feature alone. The possible reason for this is that, according to the process of data collection and human movements, eye-gaze and head movement are connected to each other: a human moves them both while exploring an environment, usually replacing a significant eye movement with a slight head movement. The combination of these modalities at a feature level allows the system to simultaneously utilize information from both sources, which is not possible to do at a decision level.

### 6.2. Imbalance of Labels

Studies related to emotion estimation often suffer from the subjectivity of labels. In our study, we had an exceptional case of subjectivity, as emotion and satisfaction can be measured only by the participants themselves and not by any third parties, such as an annotator. An additional limitation was brought by the domain of our research—tourism. As the main idea was to measure people's first impression, they could not participate twice in the same experiment, and should not be familiar with the experimental field. This means that we could not ask local citizens to participate in an experiment, which constrained the range of potential candidates to a very narrow group. These conditions resulted in a data shortage, complicating the model training stage, affecting the general performance and statistical stability of results.

Natural perception of sights as something interesting and the general conditions of an experiment led to a great imbalance of emotional labels. Some emotion groups are predefined to be almost empty a priori because these emotions can be caused by sightseeing, e.g., distressed or afraid. Even after dividing emotions into three groups (positive, negative, and neutral) unequally, we had 71% of positive samples. Many of these problems and limitations can be partially overcome by increasing the amount of data, which would make the system more stable and robust.

### 6.3. Limitation of Data Sources

To realize the proposed system, we need to collect several data sources, such as eye-gaze data, head motions, and selfie videos. However, there is a limitation in collecting such data in real-life conditions. Although eye-tracking devices are becoming smaller and cheaper year by year, it will take more time for them to be commonly and frequently used. However, head-motions can be measured because JINS MEME [51] with electro-oculography (EOG) and a six-axis inertial measurement unit (IMU) are already being sold on the market, and many people are using them in their daily life. In case of using selfie videos, we must take it into account that the emotional status on such videos may be exaggerated. One of the most possible shifts may be done from natural emotions to acted ones. If so, it is possible as future work to utilize existing databases for emotion recognition to improve the performance of the system.

The aim of our study, as our first attempt, was to reveal what kind of data are needed for estimating the emotional status and satisfaction level of tourists. Hence, we employed all kinds

of conceivable modalities in this paper, but we do not assert that all the modalities are required. Through the experiments, we have confirmed that various modalities and their fusion can be used for estimating emotional status and satisfaction level. We also found that there is no notable difference in the estimation performance at the unimodal level; however, performance can improve by combining them in several ways. This suggests that our proposed method does not rely on a specific modality or combination. In the current situation, not many tourists take a selfie video that can be used as one of the data sources in our system. However, even without videos, our proposed method can perform at a certain level. Of course, if the selfie video becomes common, like selfie photos, performance can be improved.

#### 6.4. Future Perspectives

The result of this paper provides a baseline performance for estimating the emotional status and satisfaction level of tourists. Several ways can be considered to improve performance. One is to widely explore other available modalities and their combinations. For example, foot motion and direction might be used for estimating the degree of tourist interest during sightseeing. Another is to analyze the transition of emotional status and satisfaction level of tourists during a session. Since emotional status and satisfaction level can drastically change even in the same session, this transition process might also be a valuable clue to estimate the tourists' status at the end of the session.

### 7. Conclusions

To design a more context-aware system, psychological user context should be constantly taken into account. In this study, as a typical use case, we selected the tourist domain, and aimed to estimate the emotional status and satisfaction level of tourists during a sightseeing tour based on their unconscious and natural actions, e.g., selfie videos, body movements, etc. We proposed a tourist emotion- and satisfaction-estimation method by fusing several modalities. To build the model, four kinds of modalities were employed: behavioral cues (eye and head/body movement) and audiovisual data (vocal/facial expressions). Through experiments in the real world with 22 tourists, we achieved up to 0.48 of an unweighted average recall score in the three-class emotion estimation task, and up to 1.11 of mean absolute error in the seven-level satisfaction estimation task. In addition, we found that effective features used for emotion and satisfaction estimation are different among tourists with a different cultural background. As future work, we will expand the nationalities of tourists and build a general model including consideration of cultural background.

**Author Contributions:** Conceptualization, Y.M. and D.F.; data curation, Y.M., D.F., and Y.T.; formal analysis, Y.M., D.F., and Y.T.; funding acquisition, Y.M., D.F., Y.A., K.Y., and W.M.; investigation, Y.M., D.F., and Y.T.; methodology, Y.M., D.F., and Y.T.; project administration, Y.M. and D.F.; resources, Y.A., K.Y., and W.M.; software, Y.M. and D.F.; supervision, Y.A., K.Y., and W.M.; validation, D.F.; visualization, Y.M. and D.F.; writing—original draft, Y.M. and D.F.; writing—review and editing, Y.M., D.F., Y.A., K.Y., and W.M.

**Funding:** This research was funded by the Japan Society for the Promotion of Science (JSPS) KAKENHI, grant number 16J09670 and 16H01721. It is also supported by the Ministry of Science and Higher Education of the Russian Federation, grant number 2.12795.2018/12.2, and the German Academic Exchange Service.

**Conflicts of Interest:** The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; and in the decision to publish the results.

### References

1. Alegre, J.; Garau, J. Tourist Satisfaction and Dissatisfaction. *Ann. Tour. Res.* **2010**, *37*, 52–73. [CrossRef]
2. Chen, C.F.; Chen, F.S. Experience quality, perceived value, satisfaction and behavioral intentions for heritage tourists. *Tour. Manag.* **2010**, *31*, 29–35. [CrossRef]
3. TripAdvisor. Available online: <http://www.tripadvisor.com/> (accessed on 15 October 2018).
4. Yelp. Available online: <https://www.yelp.com/> (accessed on 15 October 2018).
5. Amazon. Available online: <https://www.amazon.com/> (accessed on 15 October 2018).



6. Han, K.; Yu, D.; Tashev, I. Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. In Proceedings of the 15th Annual Conference of the International Speech Communication Association (InterSpeech '14), Singapore, 14–18 September 2014.
7. Kaya, H.; Karpov, A.A.; Salah, A.A. Robust Acoustic Emotion Recognition Based on Cascaded Normalization and Extreme Learning Machines. In Proceedings of the Advances in Neural Networks (ISNN '16), St. Petersburg, Russia, 6–8 July 2016; pp. 115–123.
8. Quack, W.Y.; Huang, D.Y.; Lin, W.; Li, H.; Dong, M. Mobile Acoustic Emotion Recognition. In Proceedings of the 2016 IEEE Region 10 Conference (TENCON '16), Singapore, 22–25 November 2016; pp. 170–174.
9. Tarnowski, P.; Kołodziej, M.; Majkowski, A.; Rak, R.J. Emotion recognition using facial expressions. *Procedia Comput. Sci.* **2017**, *108*, 1175–1184. [[CrossRef](#)]
10. Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.W.; Zafeiriou, S. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1301–1309. [[CrossRef](#)]
11. Subramaniam, A.; Patel, V.; Mishra, A.; Balasubramanian, P.; Mittal, A. Bi-modal First Impressions Recognition Using Temporally Ordered Deep Audio and Stochastic Visual Features. In Proceedings of the 14th European Conference on Computer Vision (ECCV '16) Workshops, Amsterdam, The Netherlands, 8–10 and 15–16 October 2016; pp. 337–348. [[CrossRef](#)]
12. Ringeval, F.; Sonderegger, A.; Sauer, J.; Lalanne, D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG '13), Shanghai, China, 22–26 April 2013; pp. 1–8. [[CrossRef](#)]
13. Dhall, A.; Goecke, R.; Ghosh, S.; Joshi, J.; Hoey, J.; Gedeon, T. From Individual to Group-level Emotion Recognition: EmotiW 5.0. In Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI '17), Glasgow, UK, 13–17 November 2017; ACM: New York, NY, USA, 2017; pp. 524–528. [[CrossRef](#)]
14. Sidorov, M.; Minker, W. Emotion Recognition in Real-world Conditions with Acoustic and Visual Features. In Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14), Istanbul, Turkey, 12–16 November 2014; ACM: New York, NY, USA, 2014; pp. 521–524. [[CrossRef](#)]
15. Hu, P.; Cai, D.; Wang, S.; Yao, A.; Chen, Y. Learning Supervised Scoring Ensemble for Emotion Recognition in the Wild. In Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI '17), Glasgow, UK, 13–17 November 2017; ACM: New York, NY, USA, 2017; pp. 553–560. [[CrossRef](#)]
16. Shapsough, S.; Hesham, A.; Elkhazraty, Y.; Zualkernan, I.A.; Aloul, F. Emotion recognition using mobile phones. In Proceedings of the 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (HealthCom '16), Munich, Germany, 14–16 September 2016; pp. 1–6. [[CrossRef](#)]
17. Resch, B.; Summa, A.; Sagl, G.; Zeile, P.; Exner, J.P. Urban Emotions—Geo-Semantic Emotion Extraction from Technical Sensors, Human Sensors and Crowdsourced Data. In *Progress in Location-Based Services 2014*; Springer: Cham, Switzerland, 2014; pp. 199–212. [[CrossRef](#)]
18. Petrantonakis, P.C.; Hadjileontiadis, L.J. Emotion Recognition from Brain Signals Using Hybrid Adaptive Filtering and Higher Order Crossings Analysis. *IEEE Trans. Affect. Comput.* **2010**, *1*, 81–97. [[CrossRef](#)]
19. Lin, Y.; Wang, C.; Wu, T.; Jeng, S.; Chen, J. EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '09), Taipei, Taiwan, 19–24 April 2009; pp. 489–492. [[CrossRef](#)]
20. Ringeval, F.; Eyben, F.; Kroupi, E.; Yuce, A.; Thiran, J.P.; Ebrahimi, T.; Lalanne, D.; Schuller, B.W. Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data. *Pattern Recognit. Lett.* **2015**, *66*, 22–30. [[CrossRef](#)]
21. AlHanai, T.W.; Ghassemi, M.M. Predicting Latent Narrative Mood Using Audio and Physiologic Data. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI '17), San Francisco, CA, USA, 4–9 February 2017; pp. 948–954.
22. Zheng, W.L.; Dong, B.N.; Lu, B.L. Multimodal emotion recognition using EEG and eye tracking data. In Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '14), Chicago, IL, USA, 26–30 August 2014; pp. 5040–5043. [[CrossRef](#)]
23. Soleymani, M.; Pantic, M.; Pun, T. Multimodal emotion recognition in response to videos. *IEEE Trans. Affect. Comput.* **2012**, *3*, 211–223. [[CrossRef](#)]

24. Soleymani, M.; Asghari-Esfeden, S.; Fu, Y.; Pantic, M. Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection. *IEEE Trans. Affect. Comput.* **2016**, *7*, 17–28. [CrossRef]
25. Kanjo, E.; Younis, E.M.; Sherkat, N. Towards unravelling the relationship between on-body, environmental and emotion data using sensor information fusion approach. *Inf. Fus.* **2018**, *40*, 18–31. [CrossRef]
26. Yamamoto, J.; Kawazoe, M.; Nakazawa, J.; Takashio, K.; Tokuda, H. MOLMOD: Analysis of feelings based on vital information for mood acquisition. In Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '09), Bonn, Germany, 15–18 September 2009; Volume 4.
27. Fedotov, D.; Matsuda, Y.; Takahashi, Y.; Arakawa, Y.; Yasumoto, K.; Minker, W. Towards Estimating Emotions and Satisfaction Level of Tourist based on Eye Gaze and Head Movement. In Proceedings of the 2018 IEEE International Conference on Smart Computing (SMARTCOMP), Taormina, Italy, 18–20 June 2018; pp. 399–404. [CrossRef]
28. Balandina, E.; Balandin, S.; Koucheryavy, Y.; Mouromtsev, D. IoT Use Cases in Healthcare and Tourism. In Proceedings of the 2015 IEEE 17th Conference on Business Informatics (CBI '15), Lisbon, Portugal, 13–16 July 2015; Volume 2, pp. 37–44. [CrossRef]
29. Morishita, S.; Maenaka, S.; Daichi, N.; Tamai, M.; Yasumoto, K.; Fukukura, T.; Sato, K. SakuraSensor: Quasi-Realtime Cherry-Lined Roads Detection through Participatory Video Sensing by Cars. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15), Osaka, Japan, 7–11 September 2015; pp. 695–705. [CrossRef]
30. Wu, C.H.; Lin, J.C.; Wei, W.L. Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies. *APSIPA Trans. Signal Inf. Process.* **2014**, *3*, E12. [CrossRef]
31. Matsuda, Y.; Fedotov, D.; Takahashi, Y.; Arakawa, Y.; Yasumoto, K.; Minker, W. EmoTour: Multimodal Emotion Recognition using Physiological and Audio-Visual Features. In Proceedings of the Ubiquitous Emotion Recognition with Multimodal Mobile Interfaces (UERMMI), Singapore, 8 October 2018.
32. Kassner, M.; Patera, W.; Bulling, A. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct), Seattle, WA, USA, 13–17 September 2014; pp. 1151–1160. [CrossRef]
33. Nakamura, Y.; Arakawa, Y.; Kanehira, T.; Fujiwara, M.; Yasumoto, K. SenStick: Comprehensive Sensing Platform with an Ultra Tiny All-In-One Sensor Board for IoT Research. *J. Sens.* **2017**, *2017*. [CrossRef]
34. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
35. Paltoglou, G.; Thelwall, M. Seeing Stars of Valence and Arousal in Blog Posts. *IEEE Trans. Affect. Comput.* **2013**, *4*, 116–123. [CrossRef]
36. Ying, H.; Silex, C.; Schnitzer, A.; Leonhardt, S.; Schiek, M. Automatic Step Detection in the Accelerometer Signal. In Proceedings of the 4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN '07), Aachen, Germany, 26–28 March 2007; pp. 80–85. [CrossRef]
37. Eyben, F.; Wöllmer, M.; Schuller, B.W. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia (MM '10), Firenze, Italy, 25–29 October 2010; pp. 1459–1462. [CrossRef]
38. Schuller, B.W.; Steidl, S.; Batliner, A.; Epps, J.; Eyben, F.; Ringeval, F.; Marchi, E.; Zhang, Y. The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. In Proceedings of the 15th Annual Conference of the International Speech Communication Association (InterSpeech '14), Singapore, 14–18 September 2014; pp. 427–431.
39. Baltrušaitis, T. OpenFace. 2017. Available online: <https://github.com/TadasBaltrušaitis/OpenFace> (accessed on 15 October 2018).
40. Baltrušaitis, T.; Robinson, P.; Morency, L.P. OpenFace: An open source facial behavior analysis toolkit. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV '16), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10. [CrossRef]
41. Ekman, P.; Friesen, W.V. *Manual for the Facial Action Coding System*; Consulting Psychologists Press: Sunnyvale, CA, USA, 1978.
42. Ekman, P.; Rosenberg, E.L. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*; Oxford University Press: Oxford, UK, 1997.

43. McKeown, G.; Valstar, M.; Cowie, R.; Pantic, M.; Schroder, M. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations Between a Person and a Limited Agent. *IEEE Trans. Affect. Comput.* **2012**, *3*, 5–17. [[CrossRef](#)]
44. Metallinou, A.; Yang, Z.; Lee, C.C.; Busso, C.; Carnicke, S.; Narayanan, S. The USC CreativeIT Database of Multimodal Dyadic Interactions: From Speech and Full Body Motion Capture to Continuous Emotional Annotations. *Lang. Resour. Eval.* **2016**, *50*, 497–521. [[CrossRef](#)]
45. Perepelkina, O.; Kazimirova, E.; Konstantinova, M. RAMAS: Russian Multimodal Corpus of Dyadic Interaction for Affective Computing. In Proceedings of the International Conference on Speech and Computer (SPECOM '18), Leipzig, Germany, 18–22 September 2018; pp. 501–510.
46. Fedotov, D.; Ivanko, D.; Sidorov, M.; Minker, W. Contextual Dependencies in Time-Continuous Multidimensional Affect Recognition. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC '18), Miyazaki, Japan, 7–12 May 2018; pp. 1220–1224.
47. Tieleman, T.; Hinton, G. *Lecture 6.5-RMSPProp, COURSERA: Neural Networks for Machine Learning*; Technical Report; University of Toronto: Toronto, ON, Canada, 2012.
48. Stanley, J.T.; Zhang, X.; Fung, H.H.; Isaacowitz, D.M. Cultural differences in gaze and emotion recognition: Americans contrast more than Chinese. *Emotion* **2013**, *13*, 36–46. [[CrossRef](#)] [[PubMed](#)]
49. Pragst, L.; Ultes, S.; Kraus, M.; Minker, W. Adaptive dialogue management in the kristina project for multicultural health care applications. In Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL '15), Gothenburg, Sweden, 24–26 August 2015; pp. 202–203.
50. Miehle, J.; Minker, W.; Ultes, S. What Causes the Differences in Communication Styles? A Multicultural Study on Directness and Elaborateness. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC '18), Miyazaki, Japan, 7–12 May 2018.
51. Kanoh, S.; Ichi-nohe, S.; Shioya, S.; Inoue, K.; Kawashima, R. Development of an eyewear to measure eye and body movements. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '15), Milano, Italy, 25–29 August 2015; pp. 2267–2270. [[CrossRef](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).