Conveying Emotions to Robots through Touch and Sound

Qiaoqiao Ren[1], Remko Proesmans[1], Frederick Bossuyt[2], Jan Vanfleteren[2], Francis wyffels[1], and Tony Belpaeme[1]

Faculty of Engineering and Architecture, [1]IDLab-AIRO, [2]CMST, Ghent University – imec, Technologiepark 126, 9052 Gent, Belgium
Qiaoqiao.Ren, Remko.Proesmans, Francis.wyffels, Tony.Belpaeme@ugent.be

Abstract. Human emotions can be conveyed through nuanced touch gestures. However, there is a lack of understanding of how consistently emotions can be conveyed to robots through touch. This study explores the consistency of touch-based emotional expression toward a robot by integrating tactile and auditory sensory reading of affective haptic expressions. We developed a piezoresistive pressure sensor and used a microphone to mimic touch and sound channels, respectively. In a study with 28 participants, each conveyed 10 emotions to a robot using spontaneous touch gestures. Our findings reveal a statistically significant consistency in emotion expression among participants. However, some emotions obtained low intraclass correlation values. Additionally, certain emotions with similar levels of arousal or valence did not exhibit significant differences in the way they were conveyed. We subsequently constructed a multi-modal integrating touch and audio features to decode the 10 emotions. A support vector machine (SVM) model demonstrated the highest accuracy, achieving 40% for 10 classes, with "Attention" being the most accurately conveyed emotion at a balanced accuracy of 87.65 %.

Keywords: Tactile interaction · affective computing · emotion decoding · multi-modal

1 Introduction

Touch is essential to human social interactions, serving as a vital medium for social exchange. Through touch, individuals form strong bonds and cooperative relationships, navigate social hierarchies, provide comfort and reassurance, and convey romantic interest [1]. A simple touch can communicate comfort, reassurance, empathy, or even urgency without a single word being spoken. The physical contact involved in touch encompasses a range of parameters, including pressure, duration, frequency, and location, all of which contribute to the prosocial emotional message being conveyed [1]. For instance, a light, brief touch on the shoulder may express sympathy, while a firm, prolonged grip can indicate support and solidarity [2].

Advancements in robotics and sensor technology have enabled more complex human-robot interactions. These sensors help us build robots with tactile sensitivity, allowing them to detect and respond to human emotions through touch. Tactile sensors capture nuances like pressure and texture, vital for emotion recognition [3], while auditory feedback from touch gestures enhances emotion decoding accuracy [4][5]. Several studies have explored how distinct emotions can be conveyed through various touch gestures [6]. For example, Yohanan and MacLean developed the Haptic Creature to detect touch and movement, investigating which gestures from a predefined set participants believed were appropriate for expressing nine specific emotions [3]. Previous research has attempted to decode emotions using video recordings of users touching robots but has often neglected crucial tactile elements such as the applied force [7]. Some studies instructed participants to express specific emotions through touch to see if subjective feelings could be communicated via simple vibrotactile feedback derived from touch sounds [4]. These studies primarily focused on either tactile or auditory channels in isolation, and focused on the correlations between touch gestures and emotions.

In this work, we collected data from touch and sound channels for the robot to mimic human touch perception. Specifically, we designed a tactile sensor to collect data from touch channels and additionally collect auditory signals from touch sounds. We explored the consistency of emotion expression based on touch and sound features using multi-modal methods, including sound and touch channels. We focus on the following research questions: (1) Are people inter-consistent in the way they express an emotion through touch? (2) Which emotions are distinguishable when expressed through touch? (3) Can multi-modal be used to decode different emotions?

2 Materials & Methods

2.1 Equipments

In this experiment, participants were tasked with conveying emotions to the Pepper robot, see Fig. 1. The participants were presented with definitions of the emotions to ensure they agreed with and understood the given definitions. The data collection involved the use of a custom 5-by-5 tactile sensor and a microphone to capture gesture data. Fig. 1a shows how the sensors are mounted on the Pepper robot. The following sections will detail the equipment used, the setup process, and the specific configurations implemented for data acquisition.

Tactile sensor The tactile sensor is a 5-by-5 piezoresistive pressure grid, similar to the Smart Textile we developed in [8]. As shown in Fig 2a, the sensor consists of a top and bottom electrode pattern with a piezoresistive Velostat™ sheet in between. The Velostat™ sheet was roughened by pressing it between two pieces of 80 grit sanding paper: this ensures little contact between the Velostat™ and the electrodes when no pressure is applied, and thus a small nominal sensor reading. The electrodes are made by screenprinting DuPont™ PE874™ conductive ink onto a 100 µm Bemis™ 3914 thermoplastic polyurethane (TPU) sheet. For each electrode, the printed side is facing the Velostat™ sheet. To interface the printed TPU with a printed circuit board (PCB) for readout, flex PCBs are glued to

(a) Pepper robot mounted with sensors.

(b) Participant interacting with Pepper.

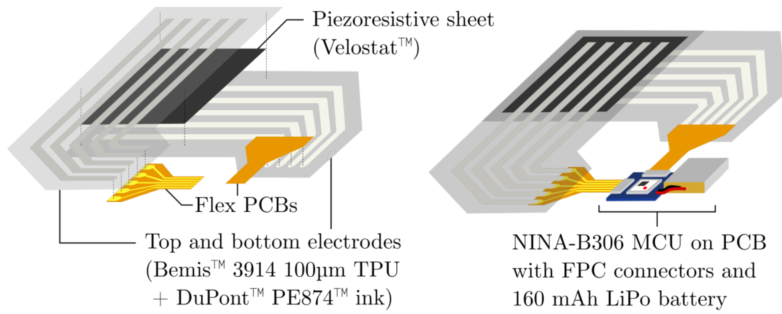
Fig. 1: Experimental setup.

the printed TPU sheet. This is achieved by adding extra conductive ink on the traces and Gorilla® Super Glue in between. Finally, extra TPU is used to cover the remaining exposed traces, to prevent oxidisation. This was omitted from Fig. 2 for clarity of the graphic. The entire structure is fused together using a Hotronix® Air Fusion IQ® heat press at 120 °C and 1.4 bar for 10 s.

The readout PCB, shown in Fig.2b, uses a NINA-B306 microcontroller unit (MCU), powered by an external 160 mAh lithium-ion polymer (LiPo) battery. Flexible printed cable (FPC) connectors are used to interface with the flex PCBs. The MCU communicates the sensor data wirelessly over Bluetooth Low Energy (BLE) at an average rate of 45 Hz. Custom Arduino board files were developed for the PCB so that it can be programmed using the Arduino IDE. We provide a guide for creating such a custom Arduino board in our GitHub repository[1].

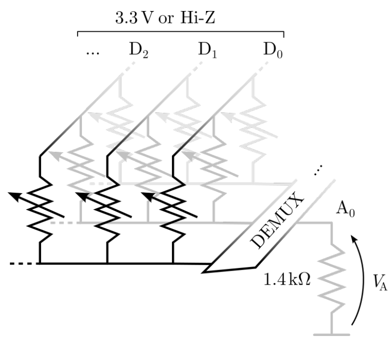
The transduction mechanism of the sensor is as follows. When pressure is applied to the sensor, the resistance of the Velostat sheet decreases. Hence, as shown in Fig.2c, the sensor can be modeled as a grid of variable resistors. Each of the traces of the top electrode is connected to a digital pin $D_{0..5}$ of the MCU. The traces of the bottom electrode run into a demultiplexer (DEMUX), of which the output is connected to an analog pin A_0 of the MCU. This pin has a 1.4 kΩ pulldown resistor. Such a low pulldown value biases the sensor as a sensitive touch sensor, rather than a pressure sensor with large dynamic range. To read a single taxel, one of the digital pins is set high (3.3 V), the others to a high impedance (Hi-Z) state. The DEMUX then routes one of the bottom traces to the A_0 pin. The voltage V_A at this pin will depend on both the pulldown resistor as well as the resistance R_x of the Velostat™ squeezed between the top trace that is driven high, and the bottom trace that is selected by the DEMUX:

[1] https://github.com/RemkoPr/airo-nrf52840-boards



(a) Layered structure. Two extra layers of TPU cover the otherwise exposed traces, omitted for clarity.

(b) Complete sensor with readout electronics.



(c) Equivalent circuit. $D_{0..5}$ and A_0 are digital and analog pins of the readout MCU.

Fig. 2: Structure of the tactile sensor.

$$V_A = \frac{1.4\,\mathrm{k\Omega}}{1.4\,\mathrm{k\Omega} + R_x} 3.3\,\mathrm{V}. \tag{1}$$

Microphone We used a Waveshare USB to audio microphone. The recordings were conducted in mono at a sample rate of 44.1 kHz. The audio stream was processed in chunks of 1024 frames per buffer. Each recorded sample, lasting 10 seconds, was saved in WAV format.

Definitions of emotions Affective states can be represented along dimensions of arousal and valence, as described by Russell's model [9]. In our study, we selected nine distinct emotions from different arousal and valence zones, plus one neutral emotion, "anger", "fear", and "disgust" fall within the high arousal and negative valence quadrant. "Happiness" and "surprise" are in the high arousal and positive valence quadrant. "sadness" and "confusion" belong to the low arousal and negative valence quadrant, while "comfort" and "calm" are from the low arousal and positive valence quadrant. "attention" is categorized as a neutral emotion. Before data collection, participants were provided with definitions of these ten emotions to ensure they understood and agreed with the interpretations. Details of each emotion definition refer to the Oxford dictionary and can be found at this GitHub repository[2].

2.2 Experimental design

To explore emotion decoding via touch-based expression and measure the consistency with which emotions and gestures are expressed across different individuals, we set up a data collection experiment. We recruited all the participants from a similar cultural background. Specifically, the group comprises 28 Chinese participants, 10 identified as female, 18 as male, and their average age is 27.8 ± 2.3 years. The data collection and study adhered to the ethics procedures of *Ghent University*, and participants gave informed consent. Participants were asked to convey each of the 10 emotions to the robot, using a gesture of their own choosing to express each emotion, with each gesture being recorded for a maximum of 10 seconds, as referred in [4]. This was repeated 3 times, so for each emotion, 3 recordings per participant were made. Participants were given time between each emotion to prepare and think about their next performance. The subsequent data collection only commenced once the participants indicated they were ready. Participants completed three rounds, with each subsequent round commencing only after all emotions or gestures had been performed, ensuring a greater variety in the data collected. After gesture data collection, we also gathered subjective feedback by asking each participant which emotions were more challenging to express.

3 Data analysis

3.1 Dataset

The dataset consists of tactile sensor and microphone recordings collected to study the conveyance of emotions through touch and audio. It includes data from 28 participants, each performing three rounds of interactions. For each of the 10 emotions, there are a total of 84 samples (28 participants for 3 rounds). Each interaction lasted 10 seconds. Tactile sensor data were sampled at a rate of 45 Hz, providing continuous feedback during the interactions. Simultaneously, audio data were recorded in 10-second WAV files.

3.2 Audio Features

Inspired by [10], the audio features extracted in this study include Mel-Frequency Cepstral Coefficients (MFCCs), Spectral Centroid, Spectral Bandwidth, Zero Crossing Rate, and Root Mean Square Energy (RMSE). MFCCs (mfcc_1 to mfcc_13) are critical in capturing the timbral qualities of touch sounds, making them effective for distinguishing different emotional tones conveyed through

[2] https://github.com/qiaoqiao2323/Convey-emotion

touch. The Spectral Centroid represents the "center of mass" of the spectrum, correlating with the perceived brightness of a sound; higher values often suggest more energetic emotions like happiness or excitement. Spectral Bandwidth measures the range of frequencies present in a sound, with broader bandwidths typically indicating more complex or intense touch sounds. The Zero Crossing Rate, which counts the number of times the audio signal changes sign, can indicate the noisiness or smoothness of a touch sound and is useful for detecting emotions like agitation or calmness. RMSE provides an overall measure of the energy present in the audio signal, with higher values potentially signifying stronger emotions such as anger or enthusiasm.

3.3 Tactile Features

Inspired by [1], tactile features extracted from each sample include Mean Pressure, Max Pressure, Pressure Variance, Pressure Gradient, Median Force, Interquartile Range Force (IQR Force), Contact Area, Rate of Pressure Change, Pressure Standard Deviation (Pressure Std), Number of Touches (Num_touches), Max Touch Duration, Min Touch Duration, and Mean Duration. Mean Pressure provides an average measure of the applied pressure, while Max Pressure identifies the peak force applied during touch interactions, both of which can reflect the intensity of emotions conveyed. Pressure Variance and Pressure Gradient offer insights into the consistency and change rate of the applied pressure, respectively, with high variability or gradient often associated with dynamic emotions like anxiety or excitement.

Median Force, representing the middle value of force applied, IQR Force measures the dispersion of the middle 50% of force values. Contact Area is crucial in identifying the nature of the touch, with larger areas potentially indicating comforting emotions, while smaller areas might suggest precision or nervousness. The Rate of Pressure Change captures how quickly the pressure varies, with high rates linked to emotions like excitement. Pressure Std reflects the fluctuation in pressure, and Numtouches quantifies the frequency of touch events within 10 seconds. Max Touch Duration, Min Touch Duration, and Mean Touch Duration describe the temporal aspects of touch, indicating how long touches are sustained, which can be correlated with different emotional expressions.

3.4 Subjective feedback

We gathered subjective feedback by asking participants which emotions were more challenging to express (multiple choice). Surprise and confusion ranked the highest, with 64.3% and 46.4% of participants choosing these emotions, respectively, out of the 10 emotions evaluated.

3.5 Consistency analysis

In order to transform feature values to a similar scale, ensuring all features contribute equally to the model, we standardized the tactile and audio features. After standardization, the numeric features of the dataset will have a mean of 0 and a standard deviation of 1. We then performed Principal Component Analysis (PCA) to reduce dimensionality and identify the key components capturing the most variance. The resulting principal component scores were integrated with the original emotion and participant identifiers. To evaluate the consistency of emotion conveyance, we calculated Intraclass Correlation Coefficients (ICC) for each emotion using the PCA scores and a two-way consistency model. As shown in Fig.3 and Table.1, all emotions demonstrated statistically significant consistency across participants, although some emotions exhibited comparatively low ICC values. The highest ICC was observed for "Attention", indicating the most consistent expression, while "Surprise" had the lowest ICC, suggesting the greatest variability in expression.

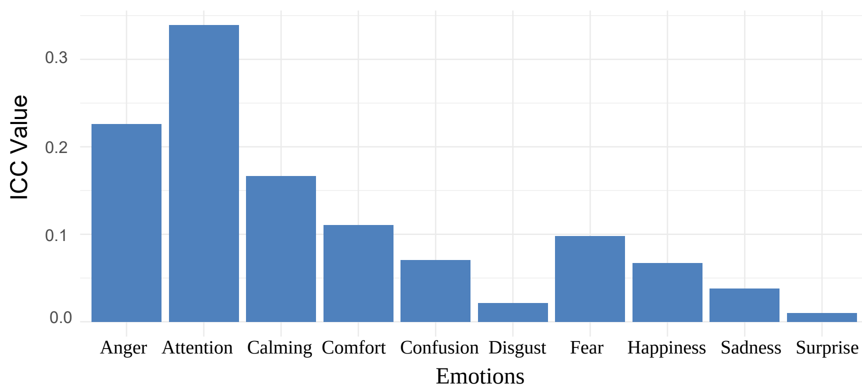


Fig. 3: ICC for emotions.

3.6 Variability analysis

The features do not follow a normal distribution. Therefore, we used permutational multivariate analysis of variance (PERMANOVA), which reveals a significant effect of emotion on the scaled combined features ($F = 14.06, p < 0.001$). Pairwise comparisons with Holm correction show that most emotion pairs differ significantly and have consistent significance after p-value adjustment. However, there is no difference between calming and sadness ($F = 3.78, p = 0.09$), comfort and sadness ($F = 4.89, p = 0.09$) based on touch and audio features, and there is no difference between confusion and comfort ($F = 4.18, p = 0.14$), confusion and sadness ($F = 1.72, p = 1.00$), disgust and sadness ($F = 3.21, p = 0.36$), disgust and surprise ($F = 1.84, p = 1.00$), happiness and surprise ($F = 2.84, p = 1.00$).

Table 1: Intraclass Correlation Coefficients among participants (** at p $<$ 0.001).

Emotions	Anger	Attention	Calming	Comfort	Confusion
ICC	0.23 **	0.34 **	0.17 **	0.11 **	0.07 **

Emotions	Disgust	Fear	Happiness	Sadness	Surprise
ICC	0.02 **	0.10 **	0.07 **	0.04 **	0.01, $p = 0.004$

Table 2: Model comparision

Models	RF	SVM	K-NN	GBM	DT	NB
Test accuracy	38.33%	40.00%	28.89%	35.00%	31.11%	28.89%

Intuitively, it is expected that “surprise” is very similar to “happiness”. One explanation could be that both of the emotions belong to the high arousal zone. Similarly, there is no significant difference between “confusion” and “sadness”, or between “disgust” and “sadness”, as they belong to the low-arousal zone.

3.7 Decoding emotions

Data split We split the training and test datasets based on participant IDs to avoid data leakage. Specifically, we randomly sampled 22 participants for the training set and 6 participants for the test set. All the extracted features, code, and visualization demos are available on GitHub[3].

Model comparison We used various machine-learning techniques, including Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Gradient Boosting Machine (GBM), Decision Tree, and Naive Bayes (NB), to decode the emotions conveyed by touch and sound due to their robustness and proficiency in handling high-dimensional data. Six models were evaluated using 10-fold cross-validation to ensure robustness and accuracy. The results of these comparisons are presented in Table 2. In addition, we explored the most significant features; we leveraged feature importance scores from Random Forest and employed Recursive Feature Elimination (RFE) with SVM, identifying the top contributing features. The top three are Max Pressure, Mean Touch Duration and RMSE, which indicate that the features from both touch and audio channel are important.

The SVM model with linear kernel demonstrated the highest overall accuracy on the test data with an accuracy of 40.00 % and a mean accuracy of 39.02 % (SD = 5.50 %) during 10-fold cross-validation. In terms of class-specific performance on the test data, for a single class, the SVM model achieved balanced accuracies [11] of 70.68 % for anger, 87.65 % for attention, 70.68 % for happiness, and 72.53 % for calming. The model also performed relatively well for confusion with a balanced accuracy of 66.98 %. However, it struggled with classes such as disgust and fear, showing balanced accuracies of 60.49 % and 59.26 %, respectively. It performs worst in sadness and surprise with a balanced accuracy of 54.01 % and 54.63 %, respectively.

Confusion matrix The confusion matrix shown in Fig. 4 indicates that attention is the most recognisable among all emotions. However, several misclassifications are evident: anger is frequently misclassified as attention, comfort is often confused with calming, and happiness is mistakenly identified as attention. Additionally, sadness is commonly confused with both comfort and calming, while surprise is often misclassified as fear and attention.

[3] https://github.com/qiaoqiao2323/Convey-emotion

These misclassifications can be attributed to the overlapping characteristics of these emotions in the context of touch and its sound. For example, the physical sensation and auditory feedback associated with anger can resemble the intense, focused signals interpreted as attention. Additionally, when participants try to convey attention to the robot, the maximum pressure applied is greater than that used for fear, happiness, and confusion. This indicates that people might not be mindful of the applied force when attempting to grab the attention of a robot. Calming and comfort are easily confused due to their similar low arousal. Happiness, conveyed through touch and sound, might be mistaken for the engaged and focused nature of attention, and it is easily confused with surprise due to their similarly high arousal. Sadness is often confused with similarly low arousal states of comfort and calming. Surprise and fear all belong to high arousal emotions, making them difficult to distinguish from each other.

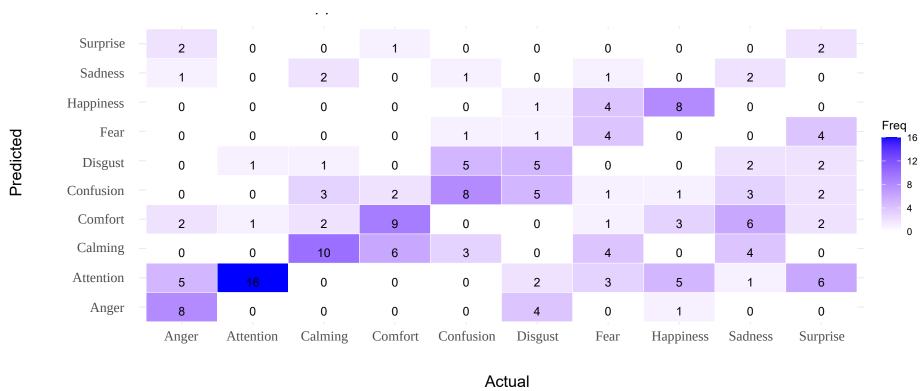


Fig. 4: SVM confusion matrix

4 Conclusion

One of our primary objectives was to determine if people consistently express the same emotions through touch. All the emotions showed significant consistency, while the ICC indicated varying degrees of consistency across different emotions. “Attention” achieved the highest consistency, suggesting that participants were more uniform in conveying this emotion. Conversely, “Surprise” had the lowest ICC, indicating considerable variability in how participants expressed this emotion through touch. The classification accuracy of the multi-modal also aligns with the ICC trend. The class “Attention” obtained the highest accuracy of 87.65 % while the class “Surprise” got the accuracy of 54.63%.

Regarding the distinguishability of emotions based on touch, consistency analyses indicated all the emotions are statistically significant consistency, although some emotions showed low ICC values. Furthermore, PERMANOVA results confirmed that most emotion pairs were significantly different. However, certain emotion pairs, like “Calming” and “Sadness” or “Confusion” and “Disgust”, did not exhibit significant differences. This suggests that these emotions share similar tactile and auditory characteristics, making them harder to distinguish.

Finally, we build a multi-modal to decode emotion; from the SVM model's confusion matrix, we can conclude that emotions are easily misclassified into categories that share similar levels of arousal or valence.

The subjective feedback collected from participants revealed that surprise and confusion were the most challenging emotions to express, with 64.3% and 46.4% of participants, respectively, indicating these emotions as difficult to express based on touch. This feedback aligns with the lower consistency scores observed for these emotions. However, this study's generalizability may be impacted by the limited sample size of 28 participants. Furthermore, the tactile sensor is currently limited to use only on the forearm.

Future research should involve extending tactile sensor coverage to multiple body parts and conveying the emotion back to the robot. This paper highlights the potential of using tactile and auditory signals for emotion recognition in human-robot interaction, which can help the robot to build an affective response during the interaction.

References

1. Matthew J Hertenstein, Dacher Keltner, Betsy App, Brittany A Bulleit, and Ariane R Jaskolka. Touch communicates distinct emotions. *Emotion*, 6(3):528, 2006.
2. Jan BF Van Erp and Alexander Toet. Social touch in human–computer interaction. *Frontiers in digital humanities*, 2:2, 2015.
3. Steven John Yohanan. *The haptic creature: Social human-robot interaction through affective touch*. PhD thesis, University of British Columbia, 2012.
4. Yulan Ju, Dingding Zheng, Danny Hynds, George Chernyshov, Kai Kunze, and Kouta Minamizawa. Haptic empathy: Conveying emotional meaning through vibrotactile feedback. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.
5. Malika Auvray, Alexandra de Lagarde, Louise P Kirsch, and Catherine Pelachaud. Social touch at a distance: Recognizing tactile gestures and their emotional content through auditory signals. In *Festival of Touch*, 2023.
6. Qiaoqiao Ren, Yuanbo Hou, and Tony Belpaeme. Low-latency classification of social haptic gestures using transformers. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 137–141, 2023.
7. Rebecca Andreasson, Beatrice Alenljung, Erik Billing, and Robert Lowe. Affective touch in human–robot interaction: conveying emotion to the nao robot. *International Journal of Social Robotics*, 10:473–491, 2018.
8. Remko Proesmans, Andreas Verleysen, Robbe Vleugels, Paula Veske, Victor-Louis De Gusseme, and Francis Wyffels. Modular piezoresistive smart textile for state estimation of cloths. *Sensors*, 22(1), 2022.
9. James A Russell et al. A cross-cultural study of a circumplex model of affect. *Journal of personality and social psychology*, 57(5):848, 1989.
10. Aisultan Shoiynbek, Kanat Kozhakhmet, Nazerke Sultanova, and Rakhima Zhumaliyeva. The robust spectral audio features for speech emotion recognition. *Appl. Math*, 13(5):867–870, 2019.
11. Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010.