SCIENCE CHINA
Information Sciences

May 2019, Vol. 62 050203:1–050203:3
https://doi.org/10.1007/s11432-018-9606-6

• LETTER •
Special Focus on Human-Robot Hybrid Intelligence

A glove-based system for object recognition via visual-tactile fusion

Bin FANG[*], Fuchun SUN, Huaping LIU, Chuanqi TAN & Di GUO

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, *China*

Received 15 August 2018/Accepted 13 September 2018/Published online 25 February 2019

Citation Fang B, Sun F C, Liu H P, et al. A glove-based system for object recognition via visual-tactile fusion. Sci China Inf Sci, 2019, 62(5): 050203, https://doi.org/10.1007/s11432-018-9606-6

Dear editor,

In our daily life, information such as tactile and visual information is widely used to recognize objects when we manipulate them. Research has shown that the human brain makes use of multisensory models of objects [1]. However, the issue of how to combine visual and tactile information to recognize the objects is challenging, because the two sensing modalities offer differing characteristics. Nowadays, the most popular method of recognizing an object is to use visual information for classification. However, practical manipulation tasks provide a great challenge for vision-based object recognition. Currently, progressive research has been performed on tactile sensing. Chitta et al. [2] studied the problem of discriminating between various types of liquid containers and their respective internal states. In addition, Schmitz et al. [3] investigated deep learning for multi-finger fusion. Liu et al. [4] exploited the intrinsic relationships between fingers and developed a joint kernel sparse coding method to combine different tactile sequences, which were captured by different fingers. Research on visual-tactile fusion object recognition is still very limited. In general, vision is suitable for dealing with color, shape, while tactile sensing is suitable for dealing with the temperature, hardness. For discerning the features of the surface material, both methods should be used. The former is usually used to deal with rough material and the latter is used to deal with finer material. Newell et al. [5] provided a detailed discussion regarding this. Recently, Gao et al. [6] proposed the joint learning method of visual images and tactile data by using the convolutional neural network (CNN). Güler et al. [7] recognized the internal state of the container focusing on transformable object with the information of visual-tactile fusion. Liu et al. [8] developed a joint group kernel sparse coding method to deal with the object recognition via visual-tactile fusion. However, in the actual robotic system, data acquisition is a relatively tedious task, and it is difficult to ensure data validity, particularly for tactile data [9].

A glove-based system is proposed for object recognition. This tactile glove is capable of jointly collecting tactile data on fingertips and the palm. It can reliably perform simultaneous tactile sensing in real time, for the purpose of collecting human hand data during fine manipulative actions. Then, the algorithms of data representation and fusion classification are deduced. Finally, we develop the visual-tactile data set for experimental verification.

System description. The tactile glove is developed to collect tactile data of human grasping. It consists of six Flexiforce pressure sensors on the front and an MCU board on the back of the glove. Five Flexiforce pressure sensors are at the five fingertips, and one is on the palm of the hand. The MCU board is used to collect and send tactile information, and the adapter board is used to switch the serial port to the USB port. The sensors' measurements are transmitted to the computer.

[*] Corresponding author (email: fangbin@tsinghua.edu.cn)

Meanwhile, a camera is used to capture images of the objects. In our system, the object is placed on the electronic turntable to capture the video. Then, we convert the video into pictures, and we can obtain the pictures of the object at different angles. We randomly extract one every 10 to 20 degrees in the pictures; hence, 30 pictures of one object are extracted as the visual dataset.

Object recognition algorithm. The covariance descriptor is used for the visual modality representation. The covariance descriptor is the integration of a variety of feature channels; it calculates their direct correlation coefficient and produces a kind of low-dimensional description of the visual features. The covariance matrix is extracted from the original image, and the feature matrix is then transformed into a 5×5 covariance descriptor.

Because grasping an object is a dynamical process and the gathered tactile information varies with time, the dynamic time warping (DTW) method is used to deal with the tactile features. The DTW algorithm uses the theory of dynamic programming to obtain the best matching path, on which the total matching distance between the two sequences reaches a minimum.

We deal with the problem of visual-tactile fusion classification by the kernel ELM (extreme learning machine) method. The details of the training process and recognition process of the visual-tactile fusion algorithm for object recognition based on kernel ELM are introduced as follows.

(1) Training process.

Input: A training set of the visual-tactile pair and its relevant label Y, Gaussian variance γ, and regularization coefficient C.

Output: W (and saving the training set of a visual-tactile pair simultaneously).

Step1: Extract visual feature. For each image of the visual train set, calculate the 5×5 covariance matrix. Calculate the covariance distance between the covariance matrix P_i and P_j of any two images, which is $d_{\rm CovD}(P_i, P_j)$. Obtain the covariance distance matrix $D_{\rm CovD}$.

Step2: Extract tactile features. Calculate DTW distance once between each tactile sequence of the tactile train set and all types of training samples, and obtain the DTW matrix $D_{\rm DTW}$.

Step3: Obtain visual feature kernel. Put $D_{\rm CovD}$ into the Gaussian kernel function to obtain

$$K_{\rm CovD} = \exp[-\gamma D^2_{\rm CovD}]. \tag{1}$$

Step4: Obtain tactile feature kernel. Put $D_{\rm DTW}$ into the Gaussian kernel function to obtain

$$K_{\rm DTW} = \exp[-\gamma D^2_{\rm DTW}]. \tag{2}$$

Step5: Visual-tactile fusion. We define the kernel of visual-tactile fusion as the product of kernel $K_{\rm CovD}$ and $K_{\rm DTW}$, which is

$$\mathrm{Kernel}(K_{\rm CovD}, K_{\rm DTW}) = K_{\rm CovD} \times K_{\rm DTW}. \tag{3}$$

Step6: Train kernel ELM classifier. Put Kernel ($K_{\rm CovD}$, $K_{\rm DTW}$) into the formula $W = (I/C + \Omega_{\rm ELM})^{-1}Y$.

(2) Recognition process.

Input: A test sample of the visual-tactile pair, W, the train set of the visual-tactile pair.

Output: Label label(x) of the test sample.

Step1: Extract visual features. Calculate the covariance distance between covariance matrix P_i and P_j of any two images, which is $d_{\rm CovD}(P_i, P_j)$. Get $1\times N$ covariance matrix $D_{\rm CovD}$ of the test sample.

Step2: Extract tactile feature. Calculate the DTW distance once between each tactile sequence of the tactile train set and all types of training samples.

Step3: Obtain visual feature kernel. Same as Eq. (1).

Step4: Obtain tactile feature kernel. Same as Eq. (2).

Step5: Visual-tactile fusion. Same as Eq. (3).

Step6: Classification. Put Kernel($K_{\rm CovD}$, $K_{\rm DTW}$) and W into $f(x) = [K(x, x_1) \ldots K(x, x_N)]W$ to determine $f(x)$. Then, according to $\mathrm{label}(x) = \arg_{i=1,\ldots,m} f_i(x)$, obtain the label($x$) of the test sample.

Experimental results. We introduce the collected data set and the experimental validation results. We selected fifteen experimental objects including a tea package, milk-tea package, coffee package, paper towel, tea box, empty water bottle, full water bottle, cylindrical box, rectangular biscuit box, soft doll, tough doll, metal bottle, paper package, tennis ball and plastic ball. In the tactile data collection process, there are six individuals involved in the collection of tactile data. Each person wore tactile gloves and grasped each experimental object five times. For each object, there are 30 pieces of tactile sequence information. The experiment randomly selects the training set and the test set from this tactile information.

Then, the proposed classification algorithm is implemented under three modes of tactile, image and visual-tactile fusion. And five different division ratios of training sets and test sets are respectively computed. The results are shown in Figure 1. The dark blue bar represents the classification results based on the visual image information. The light blue bar represents the classification results based on the tactile information. The green

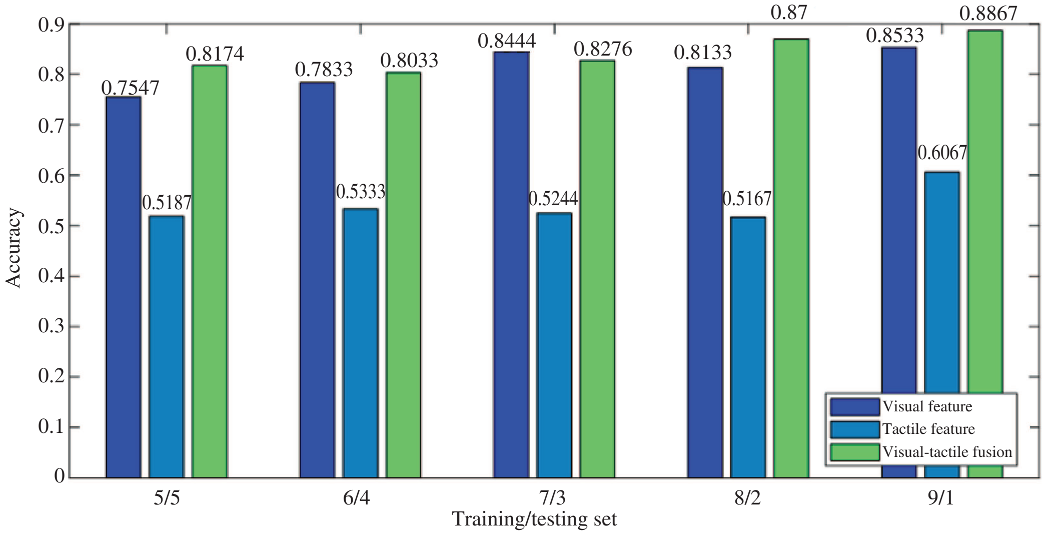

Figure 1 (Color online) Accuracy under different training/testing set division ratios.

bar represents the classification results based on the visual-tactile fusion algorithm. The histogram shows that the fusion algorithm has higher recognition accuracy than the tactile or image single-modality algorithm.

Conclusion. The designed tactile glove can be conveniently used to collect tactile information, and the visual-tactile information fusion algorithm is proposed to establish the tactile fusion object recognition system. A multi-layer time series model is used to express the tactile time series, and covariance descriptors are used to characterize the image features. The kernel ELM classification algorithm is used to fuse two kinds of modal information and to classify objects. At the same time, the tactile-visual information pair dataset, consisting of 15 objects, is established. Experimental results show that the tactile-visual fusion information classification performs significantly better than the single-modality algorithm. In the future, the results can be applied in robotic systems to improve the manipulation performance.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61503212, 61703284, U1613212), in part by National Science Foundation of China and the German Research Foundation in Project Cross Modal Learning, NSFC 61621136008/DFG TRR-169.

References

1 Lacey S, Campbell C, Sathian K. Vision and touch: multiple or multisensory representations of objects? Perception, 2007, 36: 1513–1521

2 Chitta S, Sturm J, Piccoli M, et al. Tactile sensing for mobile manipulation. IEEE Trans Robot, 2011, 27: 558–568

3 Schmitz A, Bansho Y, Noda K, et al. Tactile object recognition using deep learning and dropout. In: Proceedings of IEEE-RAS International Conference on Humanoid Robots, 2014. 1044–1050

4 Liu H P, Guo D, Sun F C. Object recognition using tactile measurements: kernel sparse coding methods. IEEE Trans Instrum Meas, 2016, 65: 656–665

5 Woods A T, Newell F N. Visual, haptic and cross-modal recognition of objects and scenes. J Physiol–Paris, 2004, 98: 147–159

6 Gao Y, Hendricks L A, Kuchenbecker K J. Deep learning for tactile understanding from visual and haptic data. 2015. ArXiv:1511.06065

7 Güler P, Bekiroglu Y, Gratal X. What's in the container? Classifying object contents from vision and touch. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2014. 3961–3968

8 Liu H P, Wu Y P, Sun F C, et al. Weakly paired multimodal fusion for object recognition. IEEE Trans Autom Sci Eng, 2018, 15: 784–795

9 Zhang W C, Sun F C, Wu H, et al. A framework for the fusion of visual and tactile modalities for improving robot perception. Sci China Inf Sci, 2017, 60: 012201