SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

January 2017, Vol. 60 012201:1–012201:12
doi: 10.1007/s11432-016-0158-2

A framework for the fusion of visual and tactile modalities for improving robot perception

Wenchang ZHANG[1,2], Fuchun SUN[1*], Hang WU[2] & Haolin YANG[1]

[1]*The State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing* 100084*, China;*
[2]*Institution of Medical Equipment, Tianjin* 300161*, China*

Received March 8, 2016; accepted June 30, 2016; published online November 22, 2016

Abstract Robots should ideally perceive objects using human-like multi-modal sensing such as vision, tactile feedback, smell, and hearing. However, the features presentations are different for each modal sensor. Moreover, the extracted feature methods for each modal are not the same. Some modal features such as vision, which presents a spatial property, are static while features such as tactile feedback, which presents temporal pattern, are dynamic. It is difficult to fuse these data at the feature level for robot perception. In this study, we propose a framework for the fusion of visual and tactile modal features, which includes the extraction of features, feature vector normalization and generation based on bag-of-system (BoS), and coding by robust multi-modal joint sparse representation (RM-JSR) and classification, thereby enabling robot perception to solve the problem of diverse modal data fusion at the feature level. Finally, comparative experiments are carried out to demonstrate the performance of this framework.

Keywords multi-modal fusion, robot perception, vision, tactile, classification

Citation 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(1): 012201, doi: 10.1007/s11432-016-0158-2

1 Introduction

Humans perceive objects synthetically using vision, tactile feedback, smell, hearing, and taste; the former two sensory inputs are most important. When humans encounter difficulty in visually discriminating objects with similar appearance, tactile sensing can help distinguish them by reasoning and synthesizing. Therefore, information of each modality and feature can improve the robustness and accuracy of classification. However, the physiological mechanism responsible for the concurrent processing of multi-modal perception information in humans is not understood clearly. Therefore, most current fusion methods for robot perception are based on probabilistic algorithms and employ Bayesian rule to combine multi-modal data. However, a critical problem in the multi-modal fusion process is that the spatial temporal patterns of feature space from multi-modal sensor data are different. In other words, perceiving an object's shape, color, and size by vision is usually static. We can classify objects based on their appearance from graphs or photos. However, tactile feedback is dynamic. We perceive an object's stiffness by the grasping process. It is, hence, very complex to fuse these two modal data.

* Corresponding author (email: fcsun@mail.tsinghua.edu.cn)

In recent years, more and more researchers have paid attention to multi-modal fusion, which is usually divided into three classes: data fusion, feature fusion, and decision fusion [1]. Fusion of various data should be in accordance with the task to be performed. Initially, multi-features or multi-tasks of vision, such as color, shape, and texture, were used in image recognition applications. These applications are usually in the same modal, either static or dynamic. Feature vectors extracted by histogram of Oriented Gradient (HOG) or scale-invariant feature transform compact (SIFT) have similar forms and represent same static character. Multimodal fusion finds wides application in human computer interaction (HCI). Multimodal systems of HCI can allow users to interact through diverse input modalities, such as speech, handwriting, hand gesture, and gaze, and can receive information from the system through output modalities, such as speech synthesis, smart graphics, and others modalities. However, these multimodal fusions are only employed at the decision-making level by histogram techniques [2], multivariate Gaussians [2], artificial neural networks (ANNs) [3,4] or hidden Markov models (HMMs) [2]. In all these systems, the probabilistic outputs of the modalities have been combined assuming conditional independence by using either the Bayesian rule or a weighted linear combination over the mode probabilities for which the weights are adaptively determined. They have not considered data properties and relationships at the feature level.

Another method that attempts to obtain better fusion results in the coding step is joint or structured sparse representation for multi-modal and multi-feature fusion. Since ref. [5] demonstrated the application of sparse coding in the receptive fields of a human being's visual cortex to extract meaningful information from images, the algorithms of sparse coding have seen rapid development in the past few years. In particular, it has led to state-of-the-art results in face recognition (FR), voice analysis, texture classification, etc. For sparse representation of multiple modalities and features, independent class including various features has within-group similarity. Furthermore, each class has a discriminative group structure for classification. Nguyen et al. [6] propose a novel multi-task multivariate (MTMV) sparse representation method for multi-sensor classification of personnel footstep recognition. The data is collected from nine sensors including acoustic, seismic, PIR and ultrasonic sensors. They add error items to sparse representation formulation in order to reduce arbitrary large noise. Zhang et al. [7] investigate the joint-structured sparsity based methods for transient acoustic signal classification with multiple measurements. Three kinds of joint structured sparse priors: same sparse code (SSC), common sparse pattern (CSP) and joint dynamic sparse (JDS) models are proposed and compared to other popular classifiers, such as nearest neighbor (NN), support vector machine (SVM), sparse representation-based classifier (SRC). They aim to employ multiple measurements from the same type of sensors rather than single channel to improve the accuracy. Yuan et al. [8] address visual multiple features classification as a multi-task joint sparse representation model for recognition. The multi-features of test images, such as color, shape, and texture, are extracted for fusion and joint sparse representation. A proximal gradient method is used to optimize the restructure dictionary. Liu and Sun [9] apply the joint sparse representation method to construct the likelihood function of particle filter tracker to enable the fusion of the color visual spectrum and thermal spectrum images for object tracking. Shekhar et al. [10] propose a multi-modal sparse representation method to constrain the observations from different modalities, which contain iris, fingerprint and face, for biometrics recognition. These modalities of information are fused by assigning more weights to the more reliable modalities, which are ranked according to the Sparsity Concentration Index (SCI). Experiments have shown that their method is robust and improves the recognition accuracy significantly. Furthermore, Rao et al. [11] propose Sparse Overlapping Group (SOG) lasso to solve problems of multi-modal grouped features that have some notion of similarity. Zhang and Levine [12] present a multi-task robust sparse representation (MRSR) model to address the fusion of multi-focus gray-level images with misregistration. It is obvious to see that joint or structured sparse representation is applied to solve the multi-task or multi-modal fusion problem by various methods.

Although various methods are proposed for multi-modal fusion at the data, feature, and decision level, most of them are not cross-modal that include dynamic and static. Moreover, there is no uniform framework for the fusion of the entire perception process from raw senor data to decision results. In this study, we propose a novel framework for the fusion of visual and tactile modal data (Figure 1) for robot perception and propose a Robust Multi-modal Joint Sparse Representation (RM-JSR) for coding.

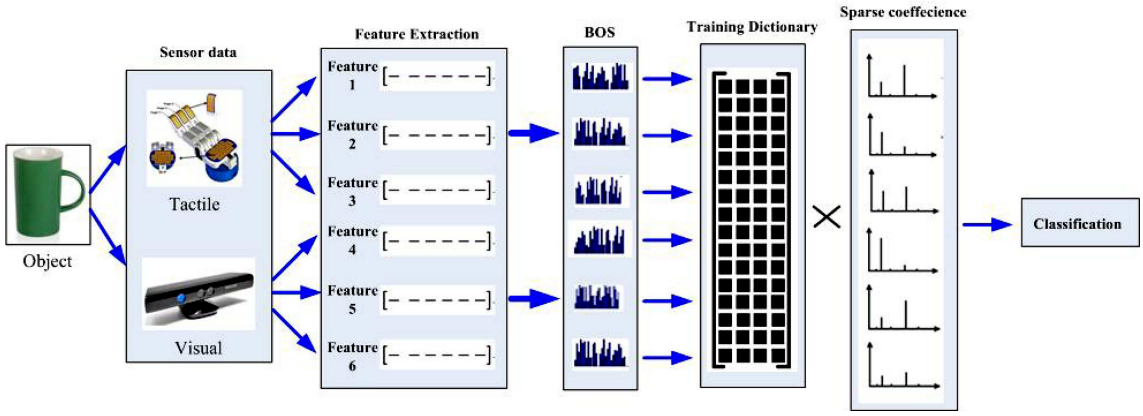

Figure 1 (Color online) Illustration of framework and pipeline of multi-modal and multi-feature fusion for robot perception. Given a query object, multi-modal raw data from various sensors are collected. Next, multi-features are extracted from visual and tactile modalities. Then, the feature vectors from different feature spaces are normalized to the same form and dimensionality by using bag-of-system (BoS). Each feature vector can be jointly represented as a linear combination product of training dictionary and sparse coefficient that have the same sparsity pattern of nonzero with different value. Finally, the dictionary is optimized and the object is classified for recognition.

The proposed framework is effective for the fusion of diverse modal data, including static and dynamic. Our framework comprises the following four steps: (1) Extraction of features from raw data obtained from various sensors, including static and dynamic modal data. (2) Normalization of feature vectors from different feature spaces based on bag-of-system (BoS). (3) Proposal of Robust Multi-modal Joint Sparse Representation (RM-JSR) for coding and classification. (4) Classification of objects by Sparse Representation based Classification (SRC).

2 Visual perception

Most object classification tasks are based on static images. There are several of features that describe images, such as SIFT [13], HOG [14], and JCD [15]. SIFT and HOG are the most widely used descriptors owing to their good performance and efficiency. After the features are extracted, the feature matrix should be encoded for classification.

Aldous et al. [16] propose bag-of-words (BoW) method for text retrieval that represents textual documents as histograms over a vocabulary of English words. Similar to the BoS approach, BoW can encode the local descriptors into a codebook that comprises codewords. Gemert et al. [17] consider codewords as characteristic representatives of the image descriptors and all codewords making up the codebook that are generated from the training image dataset. Conventional methods for codebook building include capturing the probability distribution of the local descriptors by Gaussian Mixture Model (GMM) and dividing the local descriptor space by k-means clustering. Then, the codewords are activated by encoding methods such as Hard Assignment (HA), Local Soft Assignment (LSA), Sparse Coding (SC), Locality-constrained Linear Coding (LLC) [18], and Fisher vector (FV). Finally, the feature vectors for classification can be obtained after pooling and normalization.

In conclusion, visual perception of robot includes feature extraction, encoding, and classification. The BoW method plays an important role in feature vector generation.

3 Tactile perception

Tactile sensing reading is dynamic press value in sequence that can be modeled by Linear Dynamic System (LDS) [19]. Their temporal and spatial features can be extracted from tactile sequence data that

include multiple subsequences. The LDS model is formulated as follows:

$$\begin{cases} x(t+1) = Ax(t) + Bv(t), & x(0) = x_0, \\ y(t) = Cx(t) + W(t), \end{cases} \tag{1}$$

where $x(t) \in \Re^p$ is the hidden state at time t. $A \in \Re^{n\times n}$ maps the dynamics of the hidden state, and $C \in \Re^{p\times n}$ models the hidden state of the output of the system, $W(t)$ and $Bv(t)$ are the measurements and driven by Gaussian white noise, $v(t) \sim N(0, Q)$, $w(t) \sim N(0, R)$.

Then, the dimension is reduced by principle component analysis (PCA) and use single value decomposition (SVD) to compute the observed sequence. The function is given as follows:

$$Y = U\Sigma V^{\mathrm{T}}, \tag{2}$$

where $U \in \Re^{q\times n}$, $V \in \Re^{\tau\times n}$, $\Sigma = \mathrm{diag}[\sigma_1, \ldots, \sigma_n]$ is diagonal matrix.

Then, we obtain the estimated C and X from the following function:

$$\begin{cases} \widehat{C} = U, \\ \widehat{X} = \Sigma V^{\mathrm{T}}, \end{cases} \tag{3}$$

$$\widehat{X} = [x(1), x(2), \ldots, x(\tau)], \quad A = [x(2), x(3), \ldots, x(\tau)][x(1), x(2), \ldots, x(\tau-1)]^{\dagger},$$

where † is Moore-Penrose inverse. (A, C) apply to describe the tactile feature, where A denotes the dynamic temporal feature and C represents the spatial feature.

Ellis et al. [20] and Mumtaz et al. [21] apply BoS to combine different generative models at various time resolutions through the selection of the BoS codewords, thereby resulting in superior performance. It is obvious to see that BoS can transform temporal sequence data to feature vectors easily. Therefore, for tactile data, BoS can also be used to extract features from subsequences, which can then be indicated by a histogram $h = [h_1, h_2, \ldots, h_m]^{\mathrm{T}} \in \Re^m$. Thus the tactile array can be represented by a feature vector,

$$h_{i,j} = \frac{c_{i,j}}{\sum_{j=1}^{k} c_{i,j}}, \qquad i = 1, \ldots, m; \quad j = 1, \ldots, k, \tag{4}$$

where $h_{i,j}$ is the frequency of the jth value in the ith tactile sequence. $c_{i,j}$ is the number of times of codeword j. m is the number of tactile sequences. In this way, the tactile sensing reading transforms to m-dimensional of feature vector as classifiers input [22].

The BoS method can be applied to both visual and tactile perception to encode the features, regardless of spatial or temporal data, and generate feature vectors. Therefore, BoS can normalize features from different modals into a uniform pattern in order to enable object recognition.

4 Robust multi-modal joint sparse representation (RM-JSR)

4.1 RM-JSR formulation presentation for multi-modal classification

Sparse representation has been widely used in visual classification since the past ten years. Let us consider a c class classification problem. Let matrix $X \in \Re^{d\times n}$ be an n column training feature vector of dimension d. c denotes the label of $X_j \in \Re^{d\times n_j}, j \in 1, \ldots, c$. Given a testing image feature y, the sparse representation model seeks to solve the following optimization problem:

$$\widehat{\omega} = \arg\min_{\omega} \|\omega\|_0, \qquad \text{s.t. } \|y - X\omega\|_0 \leqslant \varepsilon, \tag{5}$$

where $\|\cdot\|_0$ is ℓ_0 norm which is the sparse constraint. However, it is well-known as a NP-hard problem. Recent research has shown that l_1-norm can recover sparse representation in most cases. Therefore, it is

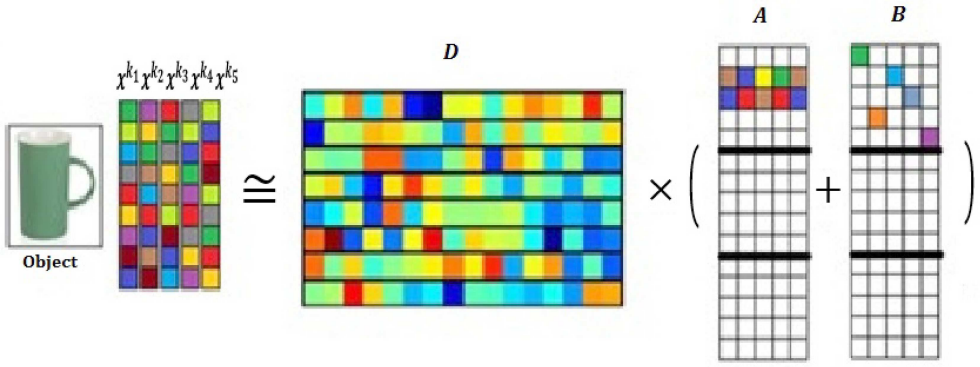

Figure 2 (Color online) x belongs to the same class but has different features. For the RM-JSR model, the sparse coefficients are the summation of A and B. A captures the share supports, while B represents unique supports from different features and classes.

transformed into a convex optimization problem that can be easily solved. Subsequently, the class label of y can be determined by the reconstruction error, as follows:

$$\hat{J} = \arg\min_{j\in 1,\ldots,c} \|y - X_j\hat{\omega}_j\|_0. \tag{6}$$

In recent years, structured sparse coding has been a trend in sparse coding (SC) research. Yuan et al. [8] propose a novel visual classification method with multitask joint sparse representation, wherein the problem is changed to a multi-task regression. It is defined as

$$\min_W \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}\|y^{kl} - X^k W^{kl}\|_2^2 + \lambda\|W\|_F, \tag{7}$$

where $X^k \in \Re^{d_k\times n}$ denote the training multi-modal features, $k = 1,\ldots,K$. $l = 1,\ldots,L$ are different samples for each feature and λ is the regularization parameter.

Sprechmann et al. [23] obtain the collaborative HiLasso model (C-HiLasso) that is very well suited for source identification and separation. It can impose the same group-sparsity pattern in all the same class samples and the different in-group sparsity patterns between samples. In addition, Jalali et al. [24] propose Dirty Block Sparse Model that defines A as the sum of shared structured atoms and B as the individual atoms. The formulation is given as follows:

$$\min_{S,B} \frac{1}{2n}\sum_{k=1}^{K}\|y^{(k)} - X^{(k)}(A^{(k)} + B^{(k)})\|_2^2 + \lambda_A\|A\|_{1,1} + \lambda_B\|B\|_{1,1}, \tag{8}$$

where k denotes k-th task. For any matrix X_m^n(m-th row $\times$ n-th column), $\|X\|_{1,1} := \sum_{m,n}|X_m^n|$ denotes the sums of absolute values; $\|X\|_{1,\infty} := \sum_m \max_n |X_m^{(n)}|$ denotes the summation of each row element's maximum absolute values. λ_A and λ_B are the regularization parameters.

In order to simplify multi-modal joint sparse representation, we impose the $l_{1,2}$-norm group-sparsity regularization on the representation matrix to jointly learn the multi-modal sparse representation. For robust representation, we propose the RM-JSR model (Figure 2) by the addition of weight vectors for different features of training data, which is given as follows:

$$\min_{D,A,B}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(\frac{1}{2}\|\omega_k x_c^{kl} - D(A_c + B_c)\|_2^2 + \lambda_1\|A_c\|_{1,2} + \lambda_2\|B_c\|_{1,1}\right), \qquad \forall c, \tag{9}$$

where $k = 1,\ldots,K$ denote each feature of the modalities index, $l = 1,\ldots,L$ is the sample times, the dictionary D is the concatenation of sub-dictionaries $D_1,\ldots,D_c$ belonging to different classes, where C is the total number of classes. For any matrix X_m^n(m-th row $\times$ n-th column), $\|X\|_{1,1} := \sum_{m,n}|X_m^n|$ denotes the sums of absolute values that induce sparsity; $\|X\|_{1,2} := \sum_m \|X_m\|_2$ denotes the sums of each row vector's l_2 norm. ω_k is the weight vector of each feature. λ_1 and λ_2 are the regularization parameters.

4.2 Optimization approach

Formulation (9) is l_1/l_2 regularization problem that can be solved by convex optimization theory. Clarke [25] proposes gradient-type methods for Optimization and non-smooth analysis. Chen et al. [26] and Schmidt et al. [27] present smooth function to approximate l_1 regularization. Figueiredo et al. [28] propose the Gradient Projection for Sparse Reconstruction (GPSR) method to solve the quadratic programming problem that is transformed by l_1/l_2 regularization. Decomposition algorithm is another effective solution. Wright et al. [29] present the Sparse Reconstruction by Separable Approximation (SpaRSA) framework and iterative methods that involve a quadratic term with diagonal Hessian plus the original sparsity-inducing regularizer. Yin et al. [30] apply Bregman Iterative Algorithms for l_1 Minimization that yields exact solutions in a finite number of steps. In recent years, Alternating Direction Methods of Multipliers (ADMM) [31] become a hot research topic for optimization solution owing to its efficiency. Complexity analysis and numerical experiments by Chi and Lange [32] show that the alternating minimization algorithm (AMA) is significantly more efficient than ADMM. Therefore, we choose AMA as the Optimization Approach (Algorithm 1). RM-JSR can be reformulated as follows:

$$\min_{A,B} \lambda_1\|A_c\|_{1,2} + \lambda_2\|B_c\|_{1,1} \quad \text{s.t.} \quad \omega_k x_c - D(A_c + B_c) = 0,\ \forall c. \tag{10}$$

We use AMA to solve Eq. (10) and reformulate as

$$\min_{A,B,P,Q} \lambda_1\|A_c\|_{1,2} + \lambda_2\|B_c\|_{1,1} \quad \text{s.t.} \quad A_c - P_c = 0, B_c - Q_c = 0, \quad \omega_k x_c - D(A_c + B_c) = 0,\ \forall c, \tag{11}$$

where $P, Q \in \Re^{K\times N}$ are alternating factors.

The augmented Lagrangian for the AMA problem is

$$\begin{aligned} L_\mu(A, B, P, Q, \mu_1, \mu_2, \mu_3) &= \lambda_1\|P\|_{1,2} + \lambda_2\|Q\|_{1,1} + \langle\mu_1, A - P\rangle + \langle\mu_2, B - Q\rangle \\ &+ \langle\mu_3, \omega_k X - D(A + B)\rangle + \frac{\mu}{2}(\|A - P\|_F^2 + \|B - Q\|_F^2 + \|\omega_k X - D(A + B)\|_F^2), \end{aligned} \tag{12}$$

where the dual variables $\mu_1, \mu_2, \mu_3 > 0$ denote a vector of Lagrange multipliers.

Algorithm 1 Sparse coding and dictionary learning by AMA

Input: Labeled training data $x_c^{kl}; k = 1, \ldots, K; l = 1, \ldots, L; c = 1, \ldots, C$, multi-feature weight vector ω_k, regularization parameters λ_1, λ_2, scalar $\rho = 1.1$

Output: $D,\ A_c,\ B_c$

1: Initializing $A^0 = 0, B^0 = 0, \mu_{1,c}^0 = 0, \mu_{2,c}^0 = 0, \mu_{3,c}^0 = 0, \mu = 1, \mu_{\max} = 10^6, n = 0$;

2: For $c = 1, \ldots, C$ do

3: While not converged do

4: Fix B_c, P_c, Q_c and update A_c

$$\begin{aligned} A_c^{n+1} &= \arg\min_{A_c} L_0(A_c, B_c, P_c^n, Q_c^n, {}_{1,c}^n, {}_{2,c}^n, {}_{3,c}^n) \\ &= (D^{\mathrm{T}}D + \mu I)^{-1}[D^{\mathrm{T}}(\omega_k X_c + \mu_{3,c}^n - DB_c^n) + \mu(P_c^{n+1} - {}_{1,c}^n)] = D^{-1}(\omega_k X_c + \mu_{3,c}^n - DB_c^n) \end{aligned}$$

5: Fix A_c, P_c, Q_c and update B_c

$$\begin{aligned} B_c^{n+1} &= \arg\min_{B_c} L_0(A_c, B_c, P_c^n, Q_c^n, \mu_{1,c}^n, \mu_{2,c}^n, \mu_{3,c}^n) \\ &= (D^{\mathrm{T}}D + \mu I)^{-1}[D^{\mathrm{T}}(\omega_k X_c + \mu_{3,c}^n - DA_c^{n+1}) + \mu(Q_c^{n+1} - {}_{2,c}^n)] = D^{-1}(\omega_k X_c + \mu_{3,c}^n - DA_c^{n+1}) \end{aligned}$$

6: Fix A_c, B_c, Q_c and update P_c

$$P_c^{n+1} = \arg\min_{P_c} L_\mu(A_c^{n+1}, B_c^{n+1}, P_c, Q_c, \mu_{1,c}^n, \mu_{2,c}^n, \mu_{3,c}^n) = \mathrm{Prox}_{\lambda_2,\Omega_{(1,2)}}(A_c^n + \mu_{1,c}^n)$$

7: Fix A_c, B_c, P_c and update Q_c

$$Q_c^{n+1} = \arg\min_{Q_c} L_\mu(A_c^{n+1}, B_c^{n+1}, P_c, Q_c, \mu_{1,c}^n, \mu_{2,c}^n, \mu_{3,c}^n) = \mathrm{Prox}_{\lambda_1,\Omega_{(1,1)}}(B_c^n + \mu_{2,c}^n)$$

8: Update Lagrange multipliers $\mu_{1,c}, \mu_{2,c}, \mu_{3,c}$

$$\begin{aligned} \mu_{1,c}^{n+1} &= \mu_{1,c}^n + \mu(A_c^{n+1} - P_c^{n+1}) \\ \mu_{2,c}^{n+1} &= \mu_{2,c}^n + \mu(B_c^{n+1} - Q_c^{n+1}) \\ \mu_{3,c}^{n+1} &= \mu_{3,c}^n + \mu(\omega_k X_c - D(A_c^{n+1} + B_c^{n+1})) \end{aligned}$$

9: Update penalty parameter $\mu = \min(\mu_{\max}, \rho\mu)$

10: Return Estimated sparse codes $A_c,\ B_c$.

The proximal maps for the ℓ_1 and ℓ_2 norms have explicit solutions, therefore, Steps 6 and 7 in Algorithm 1 can be derived as

$$\mathrm{Prox}_{\lambda_1,\Omega_{1,2}}(v_{(i,:)}) = \left(1 - \frac{\lambda_1}{\mu\|v_{(i,:)}\|_2}\right)_+ v_{(i,:)}, \tag{13}$$

$$\mathrm{Prox}_{\lambda_2,\Omega_{1,1}}(v_{(i,j)}) = \left(1 - \frac{\lambda_1}{\mu|v_{(i,j)}|_2}\right)_+ v_{(i,j)}, \tag{14}$$

where $v_{(i,j)}$ is defined as i-th row j-th column of V, and $(x)_+ = \max(x, 0)$. Finally, we get the shared sparse coefficient A_c and unique sparse coefficient B_c.

4.3 Classification

The group structured pattern enables a very simple classification. We apply sparse representation-based classification (SRC) to classify objects and only use A_c to train the classifier. Given the optimal solution $\widehat{A_j}$, the class label of y is decided based on the following criterion of minimum reconstruction error:

$$\widehat{J} = \arg\min_{j\in 1,\ldots,c} \|\omega_k y^k - D_j \widehat{A_j}\|_2. \tag{15}$$

5 Multi-modal fusion method

Estimating the weight value of each feature is crucial for multi-modal fusion. Conventionally, weight values are optimized in the dictionary learning procedure. However, it results in the addition of computational costs. Therefore, we propose a simple iteration adapted weight distribution algorithm to maximize the classification accuracy. There are mainly two factors related to the iteration efficiency, initial value and step length. We divide the samples into training data and testing data and obtain each feature's classification accuracy individually. High accuracy means more contribution. Then, we normalize and initialize the weight values, as given by the following function:

$$\omega_i = \frac{a_i}{\sum_1^k a_i}, \tag{16}$$

where a_i is classification accuracy of each feature and $i = 1, \ldots, k$ denote kth feature. ω_i is the initial weight value. The iteration adapted weight distribution algorithm is presented as Algorithm 2.

Algorithm 2 Weight value estimation

Input: step length Δd
Output: optimal ω_i
1: Initializing ω_i by formulation (16), $i = 1$;
2: For $j = i + 1, \ldots, k$ do
3: For $n = 1, \ldots, m$ do
4: while $\omega_i \geqslant 0$
5: $\omega_i = \omega_i \pm n\Delta d; \omega_j = \omega_j \mp n\Delta d$;
6: calculate accuracy a by RM-JSR framework
7: if $\Delta a \leqslant 0$
8: a=max
9: end
10: Return ω_i

6 Experiment

6.1 Experiment setup

We have not found a data-set that contains both visual and tactile data. Therefore, we build the data-set ourselves. We use Barrett hand (BH8-280) as the robotic hand platform and Kinect as visual data collection equipment. The torque output of Barrett hand was set to 2500 (programing value). We chose several-different objects for classification that look similar but have different texture and some have similar texture but appear different (Figure 3): (1) an empty beverage bottle; (2) a bottle of beverage; (3) a box of biscuits; (4) a box of toothpaste; (5) an empty water bottle; (6) a bottle of water; (7) a box

Figure 3 (Color online) Grasp objects photo.

Figure 4 (Color online) Sample photos of object.

of milk; (8) an empty milk box; (9) a can of beer; (10) an empty beer can; (11) an empty cola can; (12) a can of cola; (13) an empty paper cup; (14) a paper cup of water.

6.2 Comparative experiment

6.2.1 *Classification only by visual image*

We utilize Kinect to capture 10 photos each of 14 different objects (Figure 4). The size of each graph is 149×149 pixels. We chose five random samples of each class for training and the other five for test. Then, we extract the features by dense DSIFT (Figure 5) and HOG, respectively, and gain feature vectors by Locality-constrained Linear Coding (LLC) method. We set the size of codewords to 500 in order to get 500-dimension feature vectors. Finally, we apply SVM to classify these objects. Our algorithms are implemented in MATLAB/Windows, and run on a personal desktop computer equipped with Intel i3 CPU and 4GB RAM. The classification accuracy results by utilizing DSIFT and HOG for feature extraction are 74.2857% (52/70) (Figure 6) and 87.1429% (61/70) (Figure 7), respectively.

Analyzing the two results, we can find that the HOG feature classification accuracy is higher than DSIFT, and an unexpected error arising in the classification between cola and beer can, which have different appearances. Therefore, the classification based only on visual data is not reliable sometimes.

6.2.2 *Classification only by tactile*

We utilize the Barrett hand to grasp each of 14 objects for 10 times. The number of Barrett hand tactile sensor units is $24 \times 4 = 96$. Therefore, we obtain 96 tactile sample values (Figure 8) by fixing the frequency from beginning of grasping to stable grasping.

We use the LDS model to transform the raw tactile data and obtain the feature matrices A and C. Then, we apply two methods to classify these tactile data. First, we calculate the Martin distance of each tactile features and determine the most similar classification by k-nearest neighbors algorithm. The accuracy is 74.2857% (52/70) based on MATLAB simulation (Figure 9). Another method is the BoS model. We transform the entire time series tactile data to codewords by segmentation and cluster the codewords into 500 classification groups by K-means algorithm. Then, we generate 500 d feature vectors

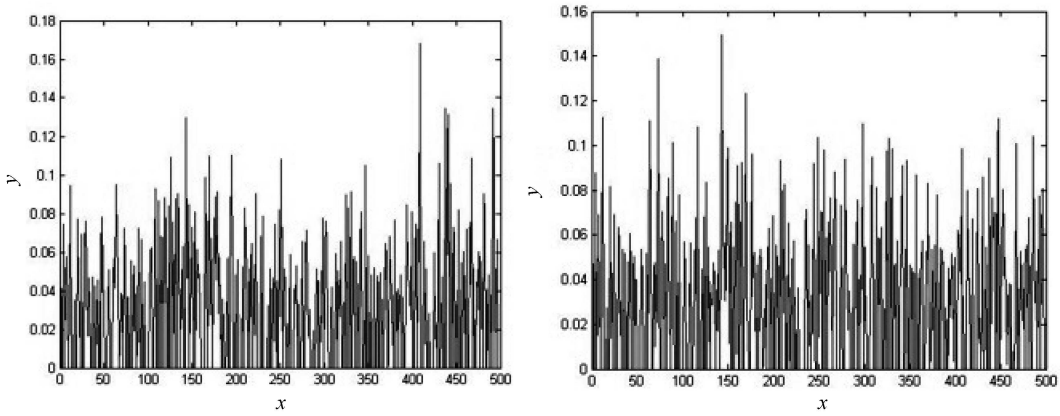

Figure 5 Two random of DSIFT feature vectors extracted from object photos; the x-axis represents code-word and y-axis represents normalization frequency.

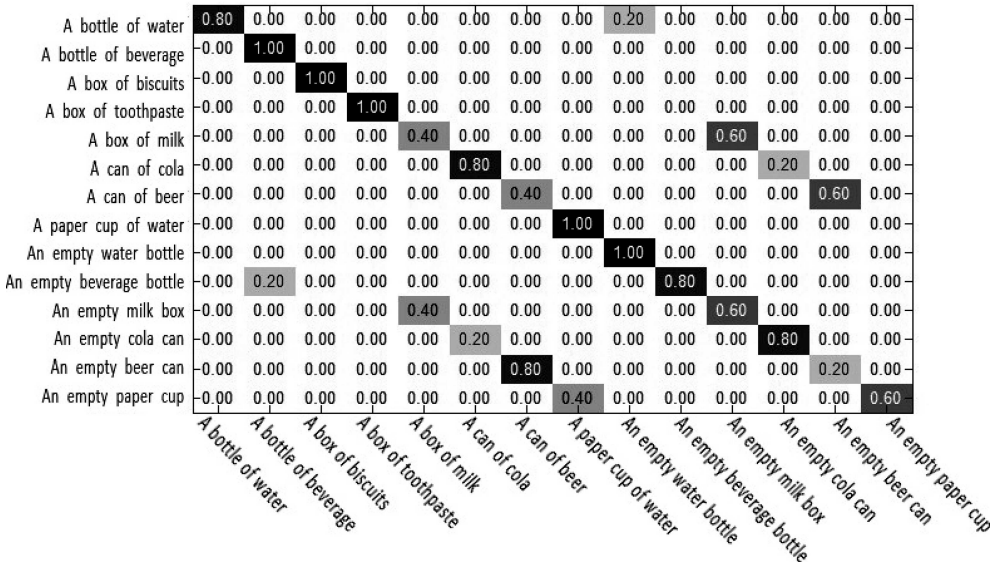

Figure 6 DSIFT feature classification accuracy by SVM is 74.2857% (52/70). There are some errors in the classification of objects with similar appearance. For example, the accuracy of 'a can of beer' is 0.4 (2/5), while error arises in the case of 'an empty beer can'. Therefore, classification only by visual data is not very accurate and credible.

by using the statistical histogram method. Fanally, these vectors are classified using SVM. The accuracy of the results obtained using the second method is 91.4286% (64/70) (Figure 10). By comparing the two tactile classification methods, it is easy to find that the BoS model play an important role in improving the classification accuracy.

6.2.3 *Classification by visual and tactile fusion*

Based on the above experiment, we can obtain visual and tactile feature vectors by BoS model; both the vectors have uniform 500 d. First, we fuse the visual and tactile feature vectors by adding proper weights and then classify them using SVM. The accuracy of the classification result is 98.5714% (69/70). Then, we apply our framework of visual and tactile fusion and utilize RM-JSR to classify these two modal data. The result is that our methods perform better and achieve 100% accuracy.

By performing these series of experiments, multi-modal fusion perception was shown to grasp more features. Moreover, its ability to discriminate improved greatly as compared to any other solo modal in term of classification accuracy. Furthermore, the proposed RM-JSR achieve better performance than SVM in terms of classification accuracy owing to its ability to capture the structure similarity between

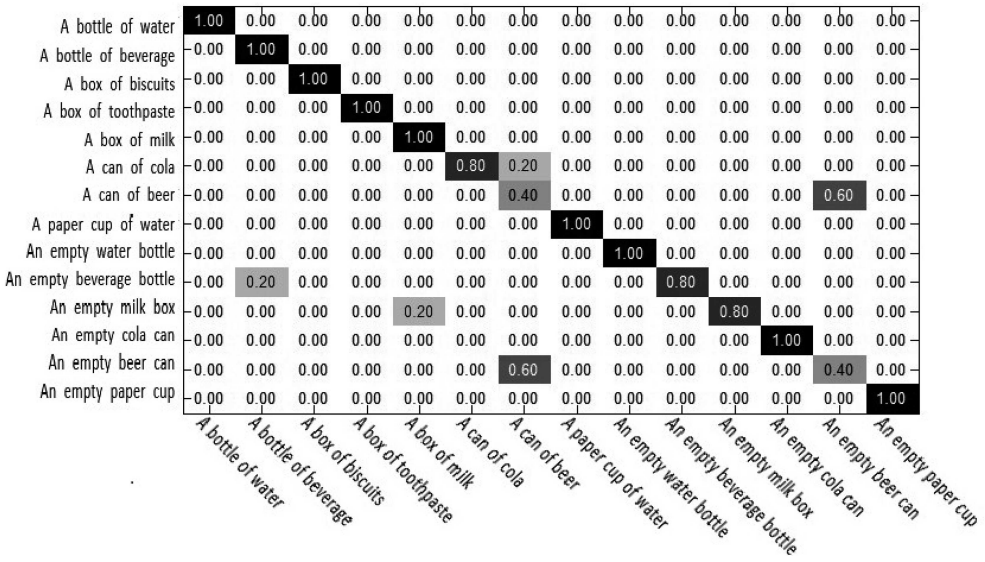

Figure 7 HOG feature classification accuracy by SVM is 87.1429% (61/70).

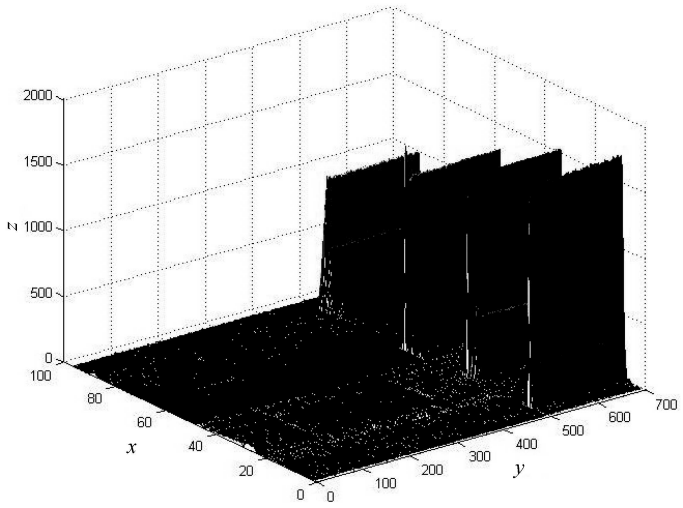

Figure 8 Tactile sensor values of Barrett hand during the task of grasping an object. The x-axis represents the number of tactile sensor units, the y-axis represents the sampling time, and the z-axis indicates the tactile values.

data in the same class and unique atoms representing in-class variation.

7 Conclusion

We present a framework for the fusion of visual and tactile data for classification purposes, which handles the static and dynamic sensing data fusion problem. Meanwhile, we develop the RM-JSR algorithm for visual and tactile classification applications. Experiments on multi-class object recognition designed by the authors show that the proposed method performs better than the single modal approach and other classification methods. In summary, we conclude that our framework is an effective method for visual and tactile fusion and can improve the classification accuracy. In the future, much more modal sensing data, such as hearing or other sensor data, will be tested and applied our framework to evaluate robot perception. Another interesting direction is to build a database that records and updates sparse representation dictionary automatically and systematically.

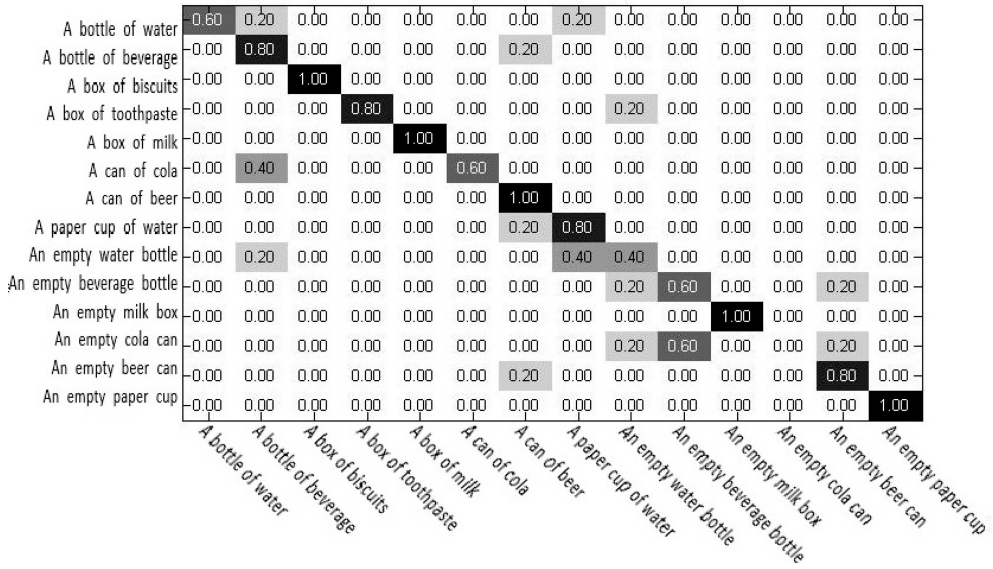

Figure 9 Classification accuracy by using Martin distance and k-nearest neighbors algorithm is 74.2857% (52/70).

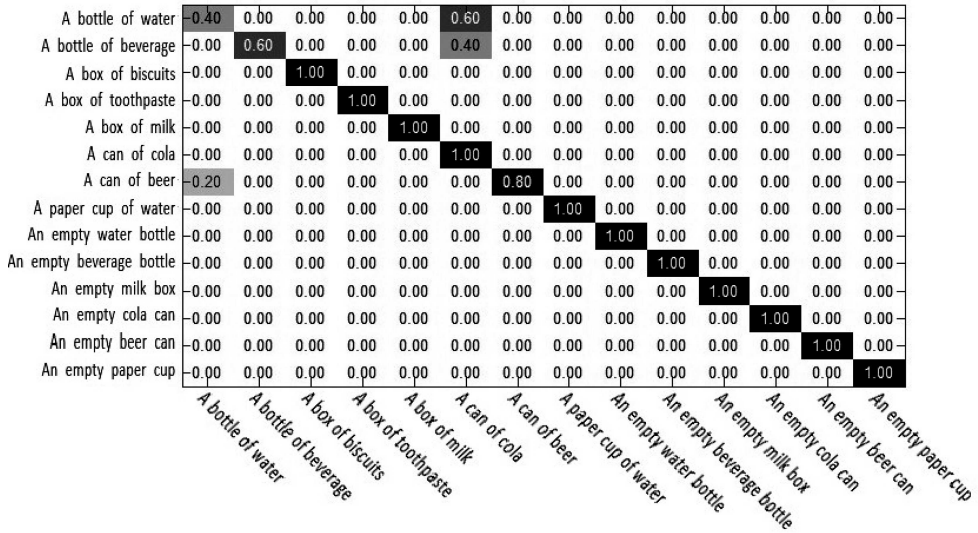

Figure 10 Classification accuracy by using BoS model and SVM is 91.4286% (64/70).

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 613278050, 61210013, 91420302, 91520201) and Academic of Military Medical Science (AMMS) Innovation Foundation (Grant No. 2015CXJJ020).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Sharma R, Pavlovic V I, Huang T S. Toward multimodal human-computer interface. Proc IEEE, 1998, 86: 853–869
2 Nock H J, Iyengar G, Neti C. Assessing face and speech consistency for monologue detection in video. In: Proceedings of the 10th ACM International Conference on Multimedia. New York: ACM, 2002. 303–306
3 Meier U, Stiefelhagen R, Yang J, et al. Towards unrestricted lip reading. Int J Pattern Recogn Artif Intell, 2000, 14: 571–585
4 Wolff G J, Prasad K V, Stork D G, et al. Lipreading by neural networks: visual processing, learning and sensory integration. In: Proceedings of Advances in Neural Information Processing Systems, Denver, 1993. 1027–1034
5 Olshausen B A, Field D J. Sparse coding with an overcomplete basis set: a strategy employed by v1? Vision Res, 1997, 37: 3311–3325

6 Nguyen N H, Nasrabadi N M, Tran T D. Robust multi-sensor classification via joint sparse representation. In: Proceedings of the 14th International Conference on Information Fusion. New York: IEEE Press, 2011. 1–8
7 Zhang H C, Zhang Y N, Nasrabadi N M, et al. Joint-structured-sparsity-based classification for multiple-measurement transient acoustic signals. IEEE Trans Syst Man Cybern-part B Cybern, 2012, 42: 1586–1598
8 Yuan X-T, Liu X B, Yan S C. Visual classification with multitask joint sparse representation. IEEE Trans Image Process, 2012, 21: 4349–4360
9 Liu H P, Sun F C. Fusion tracking in color and infrared images using joint sparse representation. Sci China Inf Sci, 2012, 55: 590–599
10 Shekhar S, Patel V M, Nasrabadi N M, et al. Joint sparse representation for robust multimodal biometrics recognition. IEEE Trans Pattern Anal Mach Intell, 2014, 36: 113–126
11 Rao N, Nowak R, Cox C, et al. Classification with the sparse group lasso. IEEE Trans Signal Process, 2016, 64: 448–463
12 Zhang Q, Levine M D. Robust multi-focus image fusion using multi-task sparse representation and spatial context. IEEE Trans Image Process, 2016, 25: 2045–2058
13 Lowe D. Distinctive image features from scale-invariant keypoints. Int J Comput Vision, 2004, 60: 91–110
14 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2005. 886–893
15 Chatzichristofis S A, Zagoris K, Boutalis Y S, et al. Accurate image retrieval based on compact composite descriptors and relevance feedback information. Int J Pattern Recogn Artif Intell, 2010, 24: 207–244
16 Aldous D, Ibragimov I, Jacod J. Exchangeability and Related Topics. Berlin: Springer, 1985. 1–198
17 van Gemert J C, Veenman C J, Smeulders A W, et al. Visual word ambiguity. IEEE Trans Pattern Anal Mach Intell, 2010, 32: 1271–1283
18 Wang J, Yang J, Yu K, et al. Locality-constrained linear coding for image classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2010. 3360–3367
19 Doretto G, Chiuso A, Wu Y N, et al. Dynamic textures. Int J Comput Vision, 2003, 51: 91–109
20 Ellis K, Coviello E, Chan A B, et al. A bag of systems representation for music auto-tagging. IEEE Trans Audio Speech Lang Process, 2013, 21: 2554–2569
21 Mumtaz A, Coviello E, Lanckriet G R G, et al. A scalable and accurate descriptor for dynamic textures using bag of system trees. IEEE Trans Pattern Anal Mach Intell, 2015, 37: 697–712
22 Ma R, Liu H P, Sun F C, et al. Linear dynamic system method for tactile object classification. Sci China Inf Sci, 2014, 57: 120205
23 Sprechmann P, Ramirez I, Sapiro G, et al. C-hilasso: a collaborative hierarchical sparse modeling framework. IEEE Trans Signal Process, 2011, 59: 4183–4198
24 Jalali A, Sanghavi S, Ruan C, et al. A dirty model for multi-task learning. In: Proceedings of Conference on Neural Information Processing Systems, Canada, 2010. 964–972
25 Clarke F H. Optimization and Nonsmooth Analysis. Hoboken: Wiley, 1990. 24–109
26 Chen X J, Zhou W J. Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth non-convex minimization. SIAM J Imag Sci, 2010, 3: 765–790
27 Schmidt M, Fung G, Rosaless R. Optimization Methods for L1 Regularization. Berlin: Springer-Verlag, 2009
28 Figueiredo M A T, Nowak R D, Wright S J. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE J Sel Topics Signal Process, 2007, 1: 586–597
29 Wright S J, Nowak R D, Figueiredo M A T. Sparse reconstruction by separable approximation. IEEE J Sel Topics Signal Process, 2009, 57: 2479–2493
30 Yin W T, Osher S, Goldfarb D, et al. Bregman iterative algorithms for l1-minimization with applications to compressed sensing. SIAM J Imag Sci, 2008, 1: 143–168
31 Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Lear, 2010, 3: 1–122
32 Chi E C, Lange K. Splitting methods for convex clustering. J Comput Graph Stat, 2015, 24: 994–1013