CSMMI: Class-Specific Maximization of Mutual Information for Action and Gesture Recognition

Jun Wan, Vassilis Athitsos, *Member, IEEE*, Pat Jangyodsuk, Hugo Jair Escalante, *Member, IEEE*, Qiuqi Ruan, *Senior Member, IEEE*, and Isabelle Guyon, *Member, IEEE*

***Abstract*— In this paper, we propose a novel approach called class-specific maximization of mutual information (CSMMI) using a submodular method, which aims at learning a compact and discriminative dictionary for each class. Unlike traditional dictionary-based algorithms, which typically learn a shared dictionary for all of the classes, we unify the intraclass and interclass mutual information (MI) into an single objective function to optimize class-specific dictionary. The objective function has two aims: 1) maximizing the MI between dictionary items within a specific class (intrinsic structure) and 2) minimizing the MI between the dictionary items in a given class and those of the other classes (extrinsic structure). We significantly reduce the computational complexity of CSMMI by introducing an novel submodular method, which is one of the important contributions of this paper. This paper also contributes a state-of-the-art end-to-end system for action and gesture recognition incorporating CSMMI, with feature extraction, learning initial dictionary per each class by sparse coding, CSMMI via submodularity, and classification based on reconstruction errors. We performed extensive experiments on synthetic data and eight benchmark data sets. Our experimental results show that CSMMI outperforms shared dictionary methods and that our end-to-end system is competitive with other state-of-the-art approaches.**

***Index Terms*— Intra-class mutual information, inter-class mutual information, class-specific dictionary, dictionary learning, Gaussian Process, sparse coding, gesture recognition, action recognition.**

Manuscript received November 4, 2013; revised March 17, 2014; accepted May 24, 2014. Date of publication June 3, 2014; date of current version June 16, 2014. This work was supported in part by the National Natural Science Foundation of China under Grant 61172128, in part by the National Key Basic Research Program of China under Grant 2012CB316304, in part by the New Century Excellent Talents in University under Grant NCET-12-0768, in part by the Program for Innovative Research Team in University, Ministry of Education of China, under Grant IRT201206, in part by the Beijing Higher Education Young Elite Teacher Project under Grant YETP0544, and in part by the National Science Foundation under Grant IIS-1055062, Grant CNS-1059235, Grant CNS-1035913, and Grant CNS-1338118. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Rafael Molina.

J. Wan and Q. Ruan are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: 09112088@bjtu.edu.cn; qqruan@center.njtu.edu.cn).

V. Athitsos and P. Jangyodsuk are with the Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: athitsos@uta.edu; pat.jangyodsuk@mavs.uta.edu).

H. J. Escalante is with the Department of Computer Science, National Institute of Astrophysics, Optics and Electronics, Puebla 72840, Mexico (e-mail: hugojair@inaoep.mx).

I. Guyon is with ChaLearn, Berkeley, CA 94708 USA (e-mail: guyon@chalearn.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2328181

I. Introduction

RECENTLY, sparse representations for human action recognition are receiving an increasing attention [1]. The theory of sparse representation aims at finding efficient and compact representations for signals. In the works of [1], three overcomplete dictionary learning frameworks were proposed using K-SVD [2]: shared dictionary (one dictionary for all classes), class-specific dictionary (one dictionary per class) and concatenated dictionary (concatenation of the class-specific dictionaries). However, K-SVD only focuses on minimizing the reconstruction error and it is not clear from [1] how to optimize the learned dictionaries. The learned dictionary obtained via K-SVD may be not compact and discriminative [3]. In this paper, we introduce a new method (named *CSMMI*) which is used to learn a compact and discriminative dictionary for each class. *CSMMI* not only discovers the latent class-specific dictionary items that best discriminates different actions, but also captures unique dictionary items for a specific class.

One of the common approaches for dictionary optimization is to use information theory [4], [5] (e.g. maximization of entropy (*ME*), maximization of mutual information (MMI)) and it shows promising results for action and gesture recognition [3], [6]. Accordingly, we adopt the MMI rule to optimize the class-specific dictionaries. However, our approach varies from the shared dictionary learning methods [3], [6]. In [6], the authors only maximize the MI for class distribution and obtain an optimal dictionary through merging of two dictionary items (see Fig. 1(a)), which can be time-consuming when the dictionary is large. We call this method *Liu-Shah*. In [3], the authors propose a Gaussian Process (GP) model for sparse representation to optimize an objective function which maximizes the MI for both appearance information and class distribution. The method [3] is referred to as *Qiu-Jiang*. However, the *Liu-Shah* and *Qiu-Jiang* methods only consider only shared dictionaries. Therefore, the optimized shared dictionary in [3] and [6] may be compact but not discriminative.

CSMMI not only considers the global information but also unifies the intra-class and inter-class MI in a single objective function. Intra-class and inter-class information is more specific and useful than the class distribution used in *Qiu-Jiang* since *CSMMI* captures discriminative dictionary items for a specific class. Our experimental results on public action and gesture recognition databases demonstrate that *CSMMI* compares favorably to the shared dictionary methods and other state-of-the-art approaches. The differences between *CSMMI* and shared dictionary methods [3], [6] are

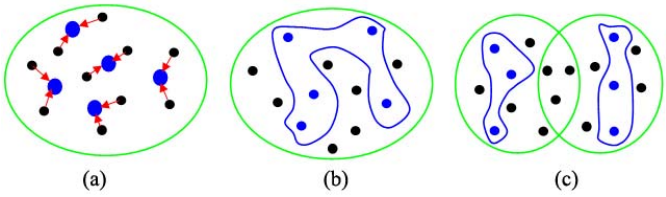


Fig. 1. (a) *Liu-Shah* [6]; (b) *Qiu-Jiang* [3], (c) *CSMMI* (our method). Each green circle denotes the region of an initial dictionary. The black points denote the initial dictionary items and the blue points represent the selected dictionary items. In the methods of *Liu-Shah* and *Qiu-Jiang*, the shared dictionary makes it difficult to distinguish which dictionary item is important to a specific class. The authors only find the dictionary items that have the minimum loss of MI. In *CSMMI*, each class have one specific dictionary and some dictionary items shared between classes can be filtered out (see the overlapped part of (c)).

shown in Fig. 1. The main contributions of this paper are:

- *CSMMI* unifies the intra-class and inter-class MI into an objective function to seek one class-specific dictionary per each class. The objective function includes two parts. The first is the intrinsic structure to keep MMI between the selected dictionary items and the rest of dictionary items in a specific class. The second is the extrinsic structure to keep minimization of mutual information (mMI) between the selected dictionary items in a specific class and the dictionary items of other classes. The aim of *CSMMI* is to select the dictionary items that are highly correlated to a specific class and less correlated to other classes.
- Because of the high computational complexity of *CSMMI*, we propose *submodularity* to calculate the class-specific dictionary. Compared with the primitive complexity $O(kC^5K^4)$, the complexity of *submodularity* is only $O(kC^2K^4)$, where K is the initial dictionary size; C is the number of classes and k ($k < K$) is the number of selected dictionary items from an initial dictionary.
- Because each class has one dictionary, we can execute class-specific dictionaries in a parallel to speed up the processing time in the recognition stage. Besides, in the initial class-specific dictionary learning stage, each action class is modeled independently of the others and hence the painful repetition of the training process when a new class is added is no longer necessary.

The rest of the paper is organized as follows. In Section II we briefly describe the related work on dictionary learning for action and gesture recognition. In Section III we present the proposed algorithm for class-specific dictionary learning in detail. *Submodularity* is proposed for reducing the complexity of *CSMMI* in Section IV. Then, in Section V, we show the experimental results and discussions. Finally, a conclusion is given in Section VI.

II. Related Work

There is a wide range of works for action and gesture recognition (see the surveys [7] and [8]). Among these methods, the dictionary-based approaches have gained widespread attention. A basic dictionary-based method is k-means that is widely used in the bag of features (BoF) model for object categorization [9]–[11] and action recognition [12], [13]. Some other methods include sparse coding-based dictionary learning [14]–[17], combined dictionary learning and classifier training [18] or information theory [3], [5], [6].

Among these dictionary-based methods, MMI clustering has revealed inspiring results, such as *Liu-Shah* [6] and *Qiu-Jiang* [3] methods. *Liu-Shah* first extracts cuboid features [19] and quantizes the cuboids using k-means to generate an initial dictionary (also called video-words). Then, MMI clustering, which groups a pair of dictionary items if they are highly correlated, is applied to find the optimal number of dictionary. *Qiu-Jiang* [3] maximizes MI for both appearance information and class distribution of dictionary items. This method first generates the initial dictionary using K-SVD [2]. Then, an objective function is defined to learn an optimal dictionary (a subset of the initial dictionary) that most reduces the entropy about the rest of the dictionary items [3]. Besides, *Qiu-Jiang* uses a Gaussian Process (GP) model to calculate the conditional entropy. However, those methods only focus on shared dictionary learning. When a new class of data is added to the action systems, shared dictionary-based methods have to repeatedly learn a new dictionary.

Besides, Mairal et al. [16] learned one dictionary for each class and used class-specfic dictionaries to achieve texture segmentation. Yang et al. [20] introduced a method to learn class-specific dictionary via fisher discrimination, which was successfully applied in face and digit recognition. From the experiments in [16] and [20], they did not evaluate their algorithm for video-based recognition tasks. Later, Guha et al. [1] proposed a random sample reconstruction method which used class-specific dictionaries to achieve action recognition. However, the authors of [1] didn't consider how to optimize class-specific dictionaries, so the learned dictionaries may be not compact and discriminative.

A. BoF Modeling and Dictionary Learning

In the BoF model, a dictionary is commonly learned by clustering (e.g. k-means, sparse coding), which has been adopted by many computer vision researchers [19], [21]. So one learned dictionary consists of a number of clustering centers and each clustering center is treated as one codeword. Then each sample vector is allowed to be approximated by one or limited codewords. For example, when the vector quantization [22] is used, each vector is assigned to one codeword that is closest to it in terms of Euclidean distance. However, this leads to a high approximation error. To reduce the approximation error, the sparsity constrain can be relaxed by allowing a few codewords to participate in the approximation process, which is the idea of sparse representation-based dictionary learning. Then, each vector can be represented by a weighted sum of a small number of codewords. Although BoF modeling method discards the spatial and temporal relationships among the codewords, it has revealed promising results [19], [21].

III. Class-Specific Sparse Representation Classification (CSSRC)

In this section, we propose a framework named CSSRC for action and gesture recognition. CSSRC includes four steps:

feature extraction and representation, learning initial class-specific dictionaries, *CSMMI* and classification. This work is inspired by [3], [23]. But [3] and [23] only focus on the shared dictionary while this work explores the relationship between intra-class and inter-class MI for video-based recognition.

A. Feature Extraction and Representation

We use four types of features in this paper. The first type is the space-time interest points (STIP) feature proposed by Laptev et al. [24]. We use STIP features to represent a video, and then histograms of oriented gradients (HOG) and histograms of optic flow (HOF) to describe each interest point. The second type is 3D enhanced motion scale invariant feature transform (EMoSIFT) [21] feature which fuses the RGB data and depth information into the feature descriptors. The third type is Histograms of 3D Joints (HOJ3D) feature [25] computed from skeleton information. The last type is shape-motion feature [26], which is used to extract shape and motion features from video sequences. For different datasets, we may use different features based on the experimental results.

B. Learning Initial Class-Specific Dictionaries

Suppose there are C classes and each class has m training samples. We first extract the feature set denoted by $Y_i = [y_1, \dots y_j, \dots y_p], y_j \in \Re^n$ (each feature descriptor has n dimensions) from m training videos in the i^{th} class, where p is the number of feature descriptors. Then, for the i^{th} class, we can obtain an initial dictionary $\Phi_i^0 = [\phi_1, \dots \phi_j, \dots \phi_K], \phi_j \in \Re^n$ with the dictionary size K over which Y_i has a sparse representation $X_{\Phi_i^0} = [x_1, \dots x_j, \dots x_p], x_j \in \Re^K$. It is formally written as the following optimization problem:

$$\min_{\Phi_i^0, X_{\Phi_i^0}} \{\|Y_i - \Phi_i^0 X_{\Phi_i^0}\|_F^2\} \quad s.t. \ \|x_j\|_0 \le T \tag{1}$$

where i denotes the i^{th} class label, $\|\cdot\|_F$ is the Frobenius norm, $\|\cdot\|_0$ is the ℓ_0 norm that counts the number of nonzero elements in a vector, and T is the sparsity parameter (i.e., the number of non-zero elements allowed). To solve Eq. 1, K-SVD [2] is considered, which is usually used to learn a dictionary for sparse coding [1], [3]. Therefore, we can get initial dictionaries $\Phi_1^0, \Phi_2^0, \dots, \Phi_C^0$ for C classes (one dictionary per class), and the concatenated dictionary is represented by $\Phi^0 = [\Phi_1^0, \Phi_2^0, \dots, \Phi_C^0]$.

C. CSMMI

Given initial dictionaries $\Phi_1^0, \Phi_2^0, \dots, \Phi_C^0$ (dictionary size $|\Phi_i^0| = K$), we aim to compress them into new dictionaries $\Phi_1^*, \Phi_2^*, \dots, \Phi_C^*$ ($|\Phi_i^*| = k$, $k < K$). For the i^{th} class with its initial dictionary Φ_i^0, we seek to learn Φ_i^* which keeps the MMI by the difference between the intra-class and inter-class MI. The objective function to find a new dictionary Φ_i^* for the i^{th} class is defined as:

$$\arg\max_{\Phi_i^*} \overbrace{I(\Phi_i^*; \Phi_i^0 \setminus \Phi_i^*)}^{intra-classMI} - \overbrace{I(\Phi_i^*; \Phi^0 \setminus \Phi_i^0)}^{inter-classMI} \tag{2}$$

where $I(\Phi_i^*; \Phi_i^0 \setminus \Phi_i^*)$ is the MI between the selected dictionary items Φ_i^* and the rest of dictionary items $\Phi_i^0 \setminus \Phi_i^*$; $I(\Phi_i^*; \Phi^0 \setminus \Phi_i^0)$ is the MI between the selected dictionary items Φ_i^* and other class dictionaries $\Phi^0 \setminus \Phi_i^0$. Our objectives are to keep MMI between the selected dictionary items and the rest of dictionary items in a specific class (see intra-class MI), and keep mMI between the selected dictionary items in a specific class and dictionary items of other classes (see inter-class MI).

It is known that maximizing the above function is NP-hard. Although the problem has been studied in the machine learning literature [23], it is only used to seek a shared dictionary. The differences of the optimization problems between the shared dictionary and class-specific dictionaries are the different objective functions and the number of dictionaries. Here, we extend the work from [23] to seek class-specific dictionary. We first initialize $\Phi_i^* = null$ (an empty matrix), then our goal is to greedily select the next best dictionary item ϕ_i that maximizes:

$$\arg\max_{\phi_i \in \Phi_i^0 \setminus \Phi_i^*} \overbrace{I(\Phi_i^* \cup \phi_i; \Phi_i^0 \setminus (\Phi_i^* \cup \phi_i)) - I(\Phi_i^*; \Phi_i^0 \setminus \Phi_i^*)}^{intra-class\ MI\ term(\tau_1)} - [\overbrace{I(\Phi_i^* \cup \phi_i; \Phi^0 \setminus \Phi_i^0) - I(\Phi_i^*; \Phi^0 \setminus \Phi_i^0)}^{inter-class\ MI\ term(\tau_2)}] \tag{3}$$

Then, we use the knowledge from information theory [27] to simplify the intra-class and inter-class MI terms.

$$\begin{aligned} \tau_1 &= H(\phi_i|\Phi_i^*) - H(\phi_i|\overline{\Phi_i^*}) \\ \tau_2 &= H(\phi_i|\Phi_i^*) - H(\phi_i|(\Phi_l \cup \Phi_i^*)) \end{aligned} \tag{4}$$

where Φ_l is the concatenated dictionary Φ^0 except Φ_i^0 of the i^{th} class, that means $\Phi_l = \Phi^0 \setminus \Phi_i^0$; $H(\cdot|\cdot)$ is the conditional entropy. The formula derivations of Eq. 4 are given in Appendix B. Hence, the objective function in Eq. 3 can be rewritten using Eq. 4,

$$\arg\max_{\phi_i \in \Phi_i^0 \setminus \Phi_i^*} H(\phi_i|\Phi') - H(\phi_i|\overline{\Phi_i^*}) \tag{5}$$

where $\Phi' = \Phi_l \cup \Phi_i^*$. Intuitively, the conditional entropy $H(\phi_i|\Phi')$ forces ϕ_i to be most different not only from the already selected dictionary items Φ_i^* but also from other class dictionaries $\Phi_l = \Phi_0 \setminus \Phi_i^0$. Besides, the greedy MI trades off this uncertainty with $-H(\phi_i|\overline{\Phi_i^*})$, which forces us to pick an item that is the least conditional entropy $H(\phi_i|\overline{\Phi_i^*})$.

To estimate the conditional entropy, a GP model for sparse representation is used. The GP model gives us a powerful property [3], [23]: given a set of dictionary items Φ and the associated sparse coefficients X_Φ, the distribution of X_ϕ given a testing dictionary item ϕ is a Gaussian whose conditional variance is given by:

$$\sigma_{\phi|\Phi}^2 = \kappa_{(\phi,\phi)} - \kappa_{(\phi,\Phi)} \kappa_{(\Phi,\Phi)}^{-1} \kappa_{(\Phi,\phi)} \tag{6}$$

where $\kappa_{(\phi,\Phi)}$ is the covariance vector with one entry for each $u \in \Phi$ with the value $\kappa_{(\phi,u)}$, and $\kappa_{(\Phi,\phi)} = \kappa_{(\phi,\Phi)}^T$. The covariance matrix is denoted by $\kappa_{(\Phi,\Phi)}$ where the entry $u, v \in \Phi$ have a value $\kappa_{(u,v)}$. Then we can evaluate $H(\phi|\Phi)$

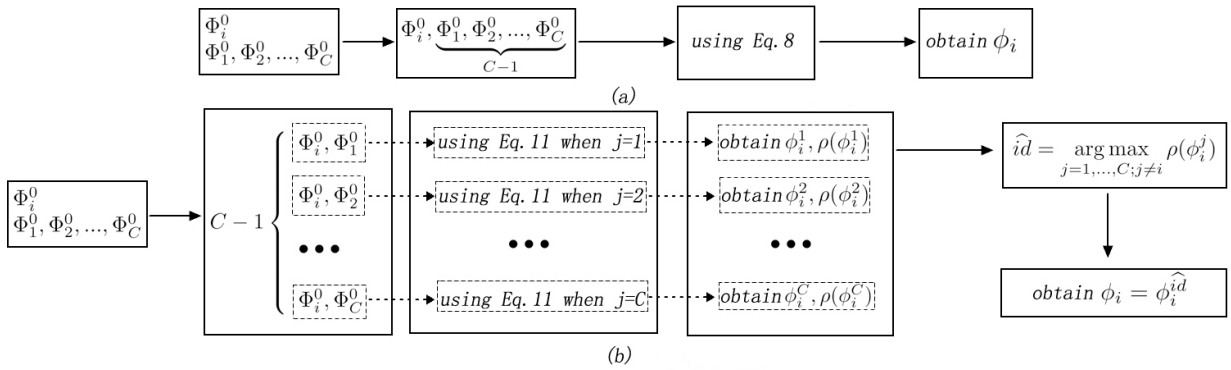


Fig. 2. It shows the flowchart to select one dictionary item ϕ_i from Φ_i^0 for the i^{th} class. (a) *CSMMI* mentioned in Section III-C; (b) *CSMMI* via *submodularity*.

as a Gaussian conditional entropy [23],

$$H(\phi|\Phi) = \frac{1}{2}\log(2\pi e\sigma^2_{\phi|\Phi}) \tag{7}$$

With the GP model, the objective function in Eq. 5 can be rewritten using Eq. 6 and Eq. 7.

$$\arg\max_{\phi_i\in\Phi_i^0\setminus\Phi_i^*} \frac{\kappa_{(\phi_i,\phi_i)} - \kappa_{(\phi_i,\Phi')}\kappa^{-1}_{(\Phi',\Phi')}\kappa_{(\Phi',\phi_i)}}{\kappa_{(\phi_i,\phi_i)} - \kappa_{(\phi_i,\overline{\Phi_i^*})}\kappa^{-1}_{(\overline{\Phi_i^*},\overline{\Phi_i^*})}\kappa_{(\overline{\Phi_i^*},\phi_i)}} \tag{8}$$

Interestingly, Eq. 8 looks like the objective function in [3, eq. 6]. However, Eq. 8 considers both the intra-class and inter-class dictionary items while [3] only considers MI between shared dictionary items. In addition, the formula derivations of Eq. 8 are given in Appendix B.

Given C classes and the initial dictionary size $|\Phi_i^0| = K$, $i = 1, 2, \ldots, C$, each iteration requires $O(C^4K^4)$ to evaluate Eq. 8. When C class dictionaries with $|\Phi_i^*| = k$ $(k < K)$ are calculated, the complexity is $O(kC^5K^4)$. It seems to be computationally infeasible for any large initial dictionary size. Therefore, we present an effective learning method to reduce the complexity by *submodularity* mentioned in Section IV.

D. Classification

A query video is represented by a collection of features $Y = [y_1, \ldots, y_j, \ldots, y_p]$, $y_j \in \Re^n$. The simple way to classify the query video is to find the smallest reconstruction error:

$$i_Y = \arg\min_{i\in[1,2,\ldots,C]} \|Y - \Phi_i^*\widehat{X_{Y_i}}\|_2^2 \tag{9}$$

where i_Y is the estimated class label and $\widehat{X_{Y_i}}$ is defined as,

$$\widehat{X_{Y_i}} = \arg\min_{X_{Y_i}} \|Y - \Phi_i^* X_{Y_i}\|_F^2 \quad s.t.\ \|x\|_0 \leq T \tag{10}$$

where $X_{Y_i} = [x_1, \ldots, x_j, \ldots, x_p]$, $x_j \in \Re^k$ is the sparse representation of Y over Φ_i^*. In our work, we use the Orthogonal Matching Pursuit (OMP) to solve Eq. 10 as in the original K-SVD paper, because it is fast and fairly accurate [2].

Owing to one dictionary per each class, we can execute Eq. 9 in a parallel way, which can speed up the processing time in the recognition stage.

IV. Reducing the Complexity of CSMMI

In Eq. 8, when we select one dictionary item for the i^{th} class, it has to use $C-1$ initial dictionaries from other classes at the same time, which leads to a high complexity $O(C^4K^4)$ in each iteration (see Fig. 2(a)). To reduce the complexity, we propose *submodularity*[1] and its flowchart is shown in Fig. 2(b). In the *submodular* method, one should note that when only one dictionary from other classes are considered to seek one dictionary item for the i^{th} class in each time, computational complexity will be significantly reduced (about $O(2^4K^4)$ in Eq. 11). Eq. 11 is defined as,

$$\phi_i^j = \arg\max_{\phi_i\in\Phi_i^0\setminus\Phi_i^*} \frac{\kappa_{(\phi_i,\phi_i)} - \kappa_{(\phi_i,\Phi^j)}\kappa^{-1}_{(\Phi^j,\Phi^j)}\kappa_{(\Phi^j,\phi_i)}}{\kappa_{(\phi_i,\phi_i)} - \kappa_{(\phi_i,\overline{\Phi_i^*})}\kappa^{-1}_{(\overline{\Phi_i^*},\overline{\Phi_i^*})}\kappa_{(\overline{\Phi_i^*},\phi_i)}} \tag{11}$$

where $\Phi^j = \Phi_j^0 \cup \Phi_i^*$; ϕ_i^j denotes a selected dictionary item from $\Phi_i^0 \setminus \Phi_i^*$ for the i^{th} class when using Φ_i^0 and Φ_j^0;

And the corresponding maximum MI is,

$$\rho(\phi_i^j) = \frac{\kappa_{(\phi_i^j,\phi_i^j)} - \kappa_{(\phi_i^j,\Phi^j)}\kappa^{-1}_{(\Phi^j,\Phi^j)}\kappa_{(\Phi^j,\phi_i^j)}}{\kappa_{(\phi_i^j,\phi_i^j)} - \kappa_{(\phi_i^j,\overline{\Phi_i^*})}\kappa^{-1}_{(\overline{\Phi_i^*},\overline{\Phi_i^*})}\kappa_{(\overline{\Phi_i^*},\phi_i^j)}} \tag{12}$$

Then, we can find the optimal index $\widehat{id}$ which has the largest MI $\rho(\phi_i^j)$.

$$\widehat{id} = \arg\max_{j=1,\ldots,C;\, j\neq i} \rho(\phi_i^j) \tag{13}$$

Finally, one best dictionary item can be selected $\phi_i = \phi_i^{\widehat{id}}$. The above mentioned process is illustrated in Figure 2(b). The complexity of *submodularity* to seek one dictionary item ϕ_i is $O((C-1)(2K)^4) \approx O(CK^4)$ (suppose initial dictionary $|\Phi_i^0| = K$). When there are C dictionaries ($|\Phi_i^*| = k$, $k < K$) to be calculated, the computational complexity is $O(kC(C-1)(2K)^4) \approx O(kC^2K^4)$. We can see that when $C \gg 2$, the complexity is significantly reduced, compared with $O(kC^5K^4)$ mentioned in Section III-C.

[1]In [23], there is also a *submodular* method, but it is different from what we are proposing in this paper. Krause et al. [23] proposed a *submodular* technique, which is a set of function F, to prove that their greedy method to seek dictionary items from a shared dictionary keeps $1 - 1/e$ of optimum, while our proposed *submodularity* tends to split the objective function Eq. 8 into $C-1$ objective functions Eq. 11 to reduce the computational complexity.

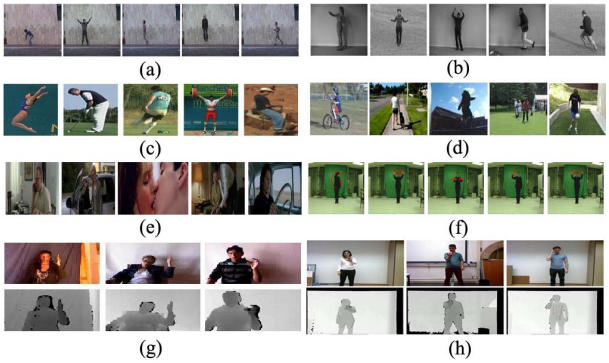


Fig. 3. Some samples from different datasets. (a) Weizmann action dataset; (b) KTH action dataset; (c) UCF sports action dataset; (d) UCF Youtube action dataset; (e) Hollywood2 action dataset; (f) Keck gesture dataset; (g) one-shot learning Chalearn gesture dataset; (h) multi-modal gesture dataset. In (g) and (h), the first row is RGB frames and the second row is the corresponding depth frames.

To elaborate on how to calculate a dictionary Φ_i^* for the i^{th} class, we first initialize C dictionaries $\Phi_1^0, \ldots, \Phi_i^0, \ldots, \Phi_C^0$. Then the covariance matrix is calculated via the associated sparse coefficients. In every $C-1$ iterations, we can select one best dictionary item ϕ_i. Lastly, we can obtain an optimal dictionary Φ_i^* via two ways. The first way is that we set the desired dictionary size $|\Phi_i^*| = k$ for each class. The second way is to use a predefined threshold ϵ to seek the optimal dictionary. *CSMMI* via *submodularity* to seek one dictionary for the i^{th} class is summarized as follows:

step 1. Calculate sparse coefficient $X_{\Phi_j^0}$ via Eq. 1 (using Y_i and Φ_j^0), where $j = 1, \ldots, C$ and Y_i is the training feature set from the i^{th} class.

step 2. Calculate covariance matrix $\kappa(\Phi_i^0, \Phi_j^0) = cov(X_{\Phi_i^0}, X_{\Phi_j^0})$, where $j = 1, \ldots, C; j \neq i$.

step 3. Calculate ϕ_i^j and $\rho(\phi_i^j)$ via Eq. 11, where $j = 1, \ldots, C; j \neq i$.

step 4. Select one best dictionary items using $\widehat{id} = \arg\max_{j=1,\ldots,C; j\neq i} \rho(\phi_i^j) \Rightarrow \phi_i = \phi_i^{\widehat{id}}$.

step 5. Continue steps 3∼4 until $\rho(\phi_i^{\widehat{id}})$ is larger than a predefined threshold ϵ or a desired dictionary size k.

In step 5, the threshold parameter ϵ is obtained empirically. In all our experiments, we simply set $\epsilon = 0.2 * \rho(\phi_i^1)$, where $\rho(\phi_i^1)$ is the MI for selecting the first dictionary item of the i^{th} class. The dictionary size of each class may be different when we use ϵ.

V. Experiments and Discussion

We evaluated the proposed method on five action recognition datasets and three gesture recognition datasets. The five action datasets are: Weizmann [28], KTH [29], UCF sports [30], UCF YouTube [31] and Hollywoods2 [32], and the three gesture datasets are: Keck [26], one-shot learning Chalearn [33] and multi-modal gesture datasets [34]. Fig. 3 shows sample frames from these datasets. The leave-one-out-cross-validation (LOOCV) is adopted for the evaluation of all the datasets unless mentioned otherwise in our experiments. For one-shot learning Chalearn gesture and multi-modal gesture datasets, the recognition performance is evaluated in predefined training/test partitions, and using the Levenshtein distance (LD) [33], also known as edit distance. Lower LD values indicate better performance. For shared dictionary-based methods, we used BoF model and a support vector machine (SVM) with intersection kernel [35] as the multi-classifier. Besides, The shared dictionary-based methods (*ME*, *Liu-Shah* and *Qiu-Jiang*) under sparse representations are described in [3] if the reader is interested in the detailed information. We set the regularized parameter $\lambda = 1$ for *Qiu-Jiang*, which agrees with the settings in [3]. All initial dictionaries (Φ_s^0 and Φ_i^0) are learned via Eq. 1 with sparsity $T = 10$.

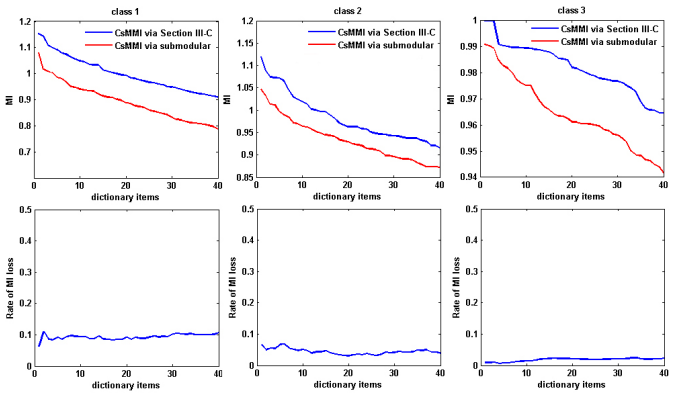


Fig. 4. It shows the MIs and rate of MI loss for each class via the direct method mentioned in Section III-C and *submodularity*.

A. Experimental Results for Synthetic Data

In this section, we first investigate the MI loss when we replace *CSMMI* mentioned in Section III-C with *submodularity*. Then, we illustrate how to eliminate shared dictionary items when some parts among different classes are overlapped.

1) MI Loss: Suppose there are three classes. Each class can be represented by a feature set $Y_i \in \Re^{512\times 1000}, i = 1, 2, 3$ which is generated based on uniform distribution (each feature descriptor has 512 dimensions). Then, we calculate the initial dictionary Φ_i^0, $i = 1, 2, 3$ for each class via Eq. 1 and each dictionary is of size 50. Finally, we can obtain the information of MI when we select each dictionary item using two methods: *CSMMI* mentioned in Section III-C and *CSMMI* via *submodularity*. To measure the lost information, we define the rate of MI loss R_l as,

$$R_l = (\rho_1(p) - \rho_2(p))/M;\ p = 1, 2, \ldots, k \qquad (14)$$

where k is the number of selected dictionary items, $\rho_1(p)$ and $\rho_2(p)$ are the MIs of the p^{th} selected dictionary item via *CSMMI* mentioned in Section III-C or *submodularity*, respectively, M is the largest MI among all of selected dictionary items and $M = max\{\rho_1(1), \rho_2(1), \ldots, \rho_1(k), \rho_2(k)\}$.

Fig. 4 shows the information of MIs via two methods to select dictionary items from Φ_i^0. In the first row, it shows the MIs when each dictionary item is selected, and the rates of MI loss R_l for each class are shown in the second row. The average rates of MI loss for three classes are 9.32%,

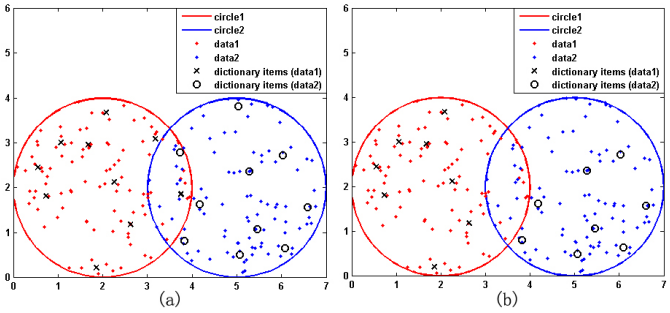


Fig. 5. (a) It shows original data and initial dictionary items (cluster centers). (b) It shows selected dictionary items from initial dictionary via *submodularity*. We can see that eight dictionary items are selected for each class and shared dictionary items are eliminated via *submodularity*.

4.41% and 1.80%, respectively. It demonstrates that *submodularity* behaves enough to keep the MI, compared with the direct method mentioned in Section III-C. Besides, *submodularity* has a low complexity and it achieves promising results in our experiments. Therefore, we will use *submodularity* to seek class-specific dictionary unless mentioned otherwise in our experiments.

2) Eliminating Shared Dictionary Items: Here, we generate two classes denoted by data1 and data2. Data1 is randomly sampled from the inner of one circle with radius 2 and center $(2, 2)$ and data2 is from the inner of one circle with radius 2 and center $(5, 2)$. Both data1 and data2 are the set of coordinates of points. Some points from both classes may be overlapped shown in Fig. 5(a). We first obtain two initial dictionaries ($K = 10$) using data1 and data2, respectively. The initial dictionary items (or cluster centers) are shown in Fig. 5(a) where two dictionary items are located in the overlapped region. Then we use *submodularity* to seek one new dictionary per class and the results are shown in Fig. 5(b). We can see that eight dictionary items are selected for each class and the shared dictionary items have been eliminated.

B. Video Representation by Sparse Coefficient Histograms

For a STIP feature set $Y \in \Re^{n\times q}$ extracted from a query video, whenever we use shared or class-specific dictionary, we can calculate sparse coefficients $X_Y = [x_1, \ldots, x_q], x_j \in \Re^k$ via Eq. 10. Therefore, the coefficient histogram is computed via $h = \frac{1}{q}\sum_{j=1}^{q} x_j$, where h is a vector with k elements. Then we normalize the coefficient histogram by ℓ_2 norm.

We learned both shared and class-specific dictionary with $|\Phi_s^0| = |\Phi_i^0| = 200$, where Φ_s^0 is the shared dictionary. Then, we applied *CSMMI* to learn the class-specific dictionaries Φ_i^* and used different shared dictionary methods (*ME*, *Liu-Shah*, *Qiu-Jiang*) to learn a shared dictionary Φ_s^*, respectively.

We selected two videos per class (one as the reference sample and the other as the test sample) as shown in the first row of Fig. 6. Those videos are represented by the coefficient histograms for both class-specific and shared dictionary methods. For shared dictionary methods, there is only one dictionary for all the classes. Therefore, a test video is represented by one histogram. As shown in Fig. 6, actions from the same class only have a few similar dominating bins in the histograms (the similar conclusion can be found in [6]) and the large reconstructive errors calculated via Eq. 9 usually occur by shared dictionary-based methods.

For *CSMMI*, the reference histograms for class-specific dictionaries are shown in the first row of Fig. 7. For a test video, we have to calculate one histogram for each class-specific dictionary. In Fig. 7, we obtain six histograms for a test video and calculate the corresponding reconstruction errors. It demonstrates that a test video belonging to a specific class not only has a relatively small reconstruction error for that specific class (see the red rectangles) but also has a similar histogram compared with the corresponding reference histogram (not only dominating bins). For example, in the second row of Fig. 7, the test video belonging to "boxing" class has a relatively small reconstruction error for the "boxing" class and its histogram is similar to the reference histogram of "boxing".

To evaluate the discrimination and compactness of the learned dictionaries, we evaluate the compactness and purity measures (under $|\Phi_s^*| = |\Phi_i^*| = 22$). The purity is the histogram of the maximum probability observing any class given a dictionary item, and the compactness is the histogram of pairwise correlation coefficients of dictionary items [3], [36]. As shown in Fig. 8, our method is the most compactness and the second most purity, compared with *ME*, *Liu-Shah*, and *Qiu-Jiang*. We note that because there is one dictionary per each class in our method, we calculate the average purity and compactness of *CSMMI* in Fig. 8.

C. Weizmann Action Dataset

It consists of ninety low-resolution (180×144 pixels) video sequences of nine subjects, each performing 10 natural actions: bend, jack (jumping jack), jump (jump forward), pjump (jump in place), run, side (gallop sideways), skip, walk, wave1 (wave one hand) and wave2 (wave two hands).

We first extracted STIP features [24] and initialized $|\Phi_s^0| = |\Phi_i^0| = 500$. The average accuracies using initial dictionaries (Φ_s^0 and Φ_i^0) are 83.3% for shared dictionary methods and 87.8% for the class-specific dictionary method.

Then we applied dictionary-based methods with different dictionary sizes k and recognition rates are shown in Fig. 9(a). It shows that *CSMMI* outperforms other methods (*ME*, *Liu-Shah* and *Qiu-Jiang*). Interestingly, when $|\Phi_i^*| \in [140, 275]$, we can obtain 100% recognition rate. Compared with other methods, their best recognition rates are: 93.44% for *ME*, 92.33% for *Liu-Shah* and 92.33%[2] for *Qiu-Jiang*. Besides, when $|\Phi_i^*| > 275$, the recognition rates slowly decrease probably caused by over-clustering. For more comparisons, we set $|\Phi_s^0| = 1000$ and performed shared dictionary-based methods with different dictionary sizes $|\Phi_s^*| = k$ $(50 \le k \le 500)$. The corresponding best recognition rates are 91.11% for *ME*, 93.33% for *Liu-Shah* and 95.56% for *Qiu-Jiang*. The results reveal that our method has a

[2]For sake of a fair comparison, we used the STIP feature for *ME*, *Liu-Shah*, *Qiu-Jiang* and *CSMMI* in our experiments. While in [3], the authors used complex features (shape and motion) to obtain 100% on Weizmann Action Dataset. This is why the recognition rate (92.33%) reported in our paper is different from that reported in [3].

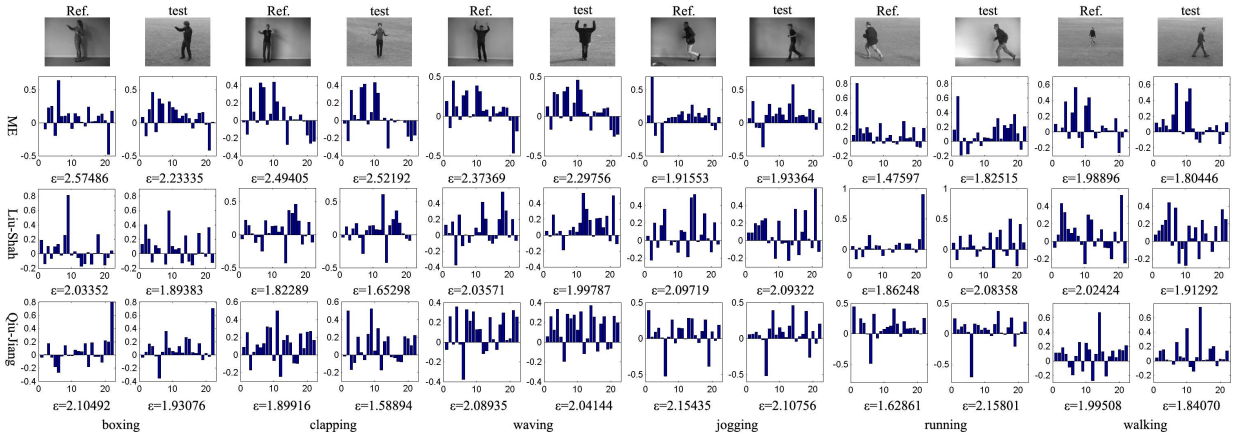


Fig. 6. The coefficient histograms of the shared dictionaries ($|\Phi_s^*| = 22$) on KTH dataset by different methods (*ME*, *Liu-Shah* and *Qiu-Jiang*). It shows that actions from the same class only have a few similar dominating bins in the histograms and large reconstructive errors usually occur in these three methods.

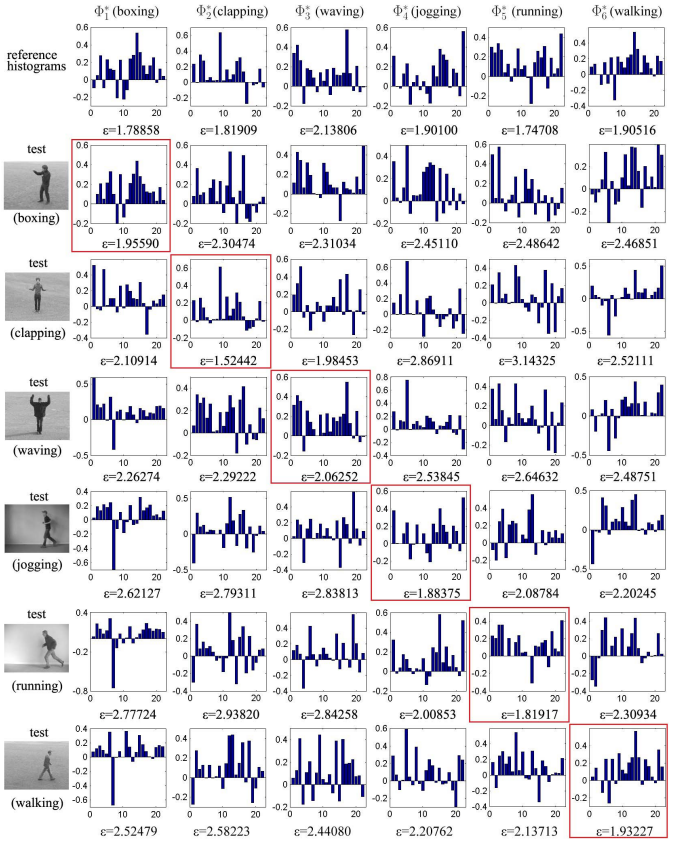


Fig. 7. The coefficient histograms of class-specific dictionaries ($|\Phi_i^*| = 22$) on KTH dataset. The first row is the reference histograms for six action classes. The rest of rows are the test video clips. It demonstrates that a test video belonging to a specific class usually has a relatively small reconstruction error for that specific class (see the red rectangles) and has a similar histogram compared with the corresponding reference histogram.

higher recognition rate than shared dictionary-based methods, even though a small dictionary size is used by *CSMMI*.

For more comparisons, the proposed approach is compared with a number of existing approaches, all of which use the LOOCV scheme to evaluate their respective algorithms. The performances are shown in Table I. Our approach achieves

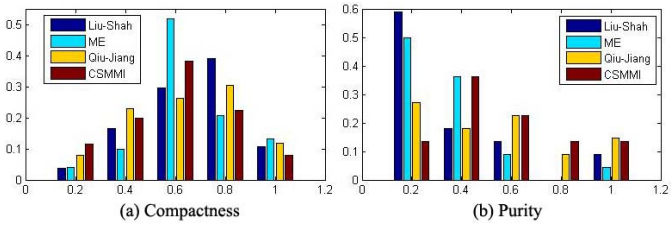


Fig. 8. Purity and compactness comparisons with dictionary size 22 ($|\Phi_s^*| = |\Phi_i^*| = 22$) on KTH action dataset. At the right-most bin of both (a) and (b), a discriminate and compact dictionary should exhibit high purity and small compactness [3].

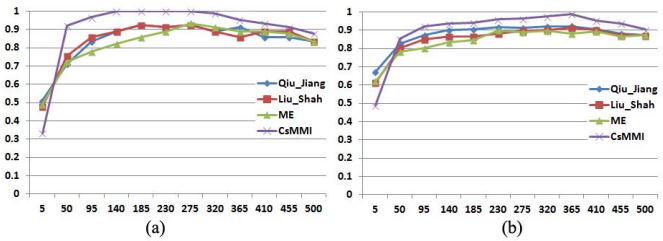


Fig. 9. The performance comparison of different methods with different dictionary sizes k. (a) Weizmann action dataset; (b) KTH action dataset.

remarkable recognition rates and outperforms most of the recent methods reported in the literature. Besides, with such basic features (STIP), we obtain 100% average recognition accuracy with only $|\Phi_i^*| = 140$. Though we use the simple STIP feature, our accuracy is comparable to the performance reported in [37] which uses complex features (silhouette-based features). We know that silhouette features are hard to obtain in complex environments, such as dynamic background, temporal scale or rotation.

D. KTH Action Dataset

This dataset consists of six actions: walking, jogging, running, boxing, waving, and clapping. They are performed by 25 actors under four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. Six hundred video sequences (160×120 pixels)

TABLE I
COMPARISON ON THE WEIZMANN ACTION DATASET

Papers	Methods	Dictionary Size	Average Accuracy
[38]	Space-time shape	—	97.83%
[39]	Multiple instance learning +kinematic feature	—	95.75%
[37]	Sparse linear approximation + feature covariance matrices	—	100%
[26]	prototype trees	—	100%
[22]	pLSA+cuboid	1200	90%
[1]	Concatenated dictionary +LMP	256	98.9%
[40]	Self-Similarities	—	95.3%
our method	*CSMMI*+STIP	140	**100%**

TABLE II
COMPARISON ON THE KTH ACTION DATASET

Papers	Methods	Dictionary Size	Average Accuracy
[29]	non-linear SVM+STIP	4000	91.8%
[39]	multiple instance learning +kinematic feature	—	87.7%
[41]	probabilistic spatiotemporal voting	—	88.0%
[37]	sparse linear approximation +Feature Covariance Matrices	—	97.4%
[26]	prototype trees	—	95.77%
[42]	Independent subspace analysis	—	93.9%
our method	*CSMMI*+STIP	365	**98.83%**

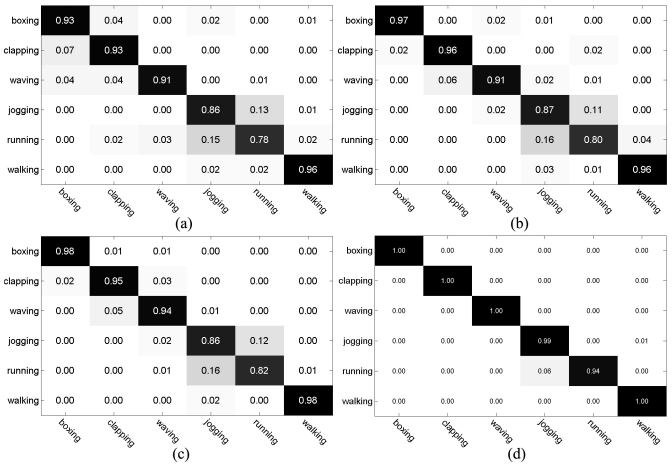


Fig. 10. Confusion matrix for KTH dataset using different methods. The average recognition rates are given: (a) *ME* (89.56%); (b) *Liu-Shah* (91.17%); (c) *Qiu-Jiang* (92.17%); (d) *CSMMI* (98.83%).

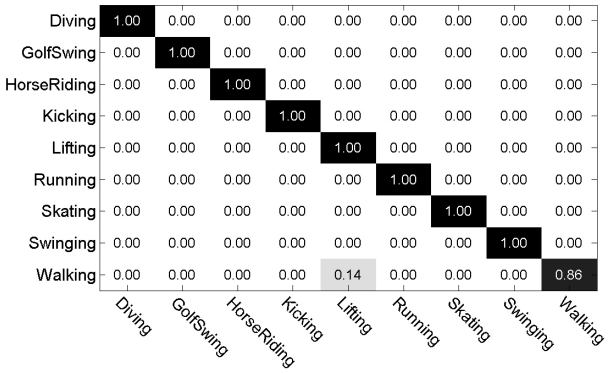


Fig. 11. Confusion matrix for UCF sports action dataset via *CSMMI*.

are recorded. This dataset is challenging due to the significant intraclass variations in terms of speed, spatial, and temporal scale, clothing, and movement variations.

We extracted STIP features [24] and initialized $|\Phi_s^0| = |\Phi_i^0| = 500$. The average accuracies using initial dictionaries (Φ_s^0 and Φ_i^0) are 87.18% for shared dictionary methods and 90.33% for the class-specific dictionary method.

The recognition rates with different dictionary sizes are shown in Fig. 9(b) which reveals that *CSMMI* outperforms the rest. More specifically, the best recognition rate for *CSMMI* is 98.83%, compared to 92.17% for *Qiu-Jiang*, 91.17% for *Liu-Shah* and 89.56% for *ME*. The confusion matrices corresponding to the best recognition results are presented in Fig. 10. We can see that the main error occurs between "running" and "jogging" in shared dictionary-based methods. However, *CSMMI* can reduce this confusion. That is because the aims of our objective functions are to select discriminative dictionary items which are highly correlated to a specific class (intra-class MI) and less correlated to other classes (inter-class MI).

Table II compares our results with some published papers which have used this dataset. Our method achieves the highest accuracy. We note that because the published papers in Table II use different experimental setups, the comparison is not totally reliable. However, the comparisons are still informative and encouraging for the researchers.

E. UCF Sports Action Dataset

The UCF sports dataset is considered to be one of the most challenging datasets in the field of action recognition. This dataset contains 150 action sequences collected from various sports videos which are typically featured on broadcast television channels such as the BBC and ESPN. The collection represents a natural pool of actions featured in a wide range of scenes and viewpoints. It also exhibits occlusion, cluttered background, variations in illumination and scale. The nine actions are: diving, golf swinging, kicking, lifting, horse riding, running, skating, swinging and walking.

We extracted STIP features [24] and initialized $|\Phi_s^0| = |\Phi_i^0| = 500$. The average accuracies using the initial dictionaries (Φ_s^0 and Φ_i^0) are 79.67% for shared dictionary methods and 83.33% for the class-specific dictionary method. Then, we can learn the class-specific dictionaries ($\Phi_i^*, i = 1, \ldots, C$) by *CSMMI* (using ε) and the corresponding average dictionary size is 469. We obtain 98% recognition rate and the corresponding confusion matrix is shown in Fig. 11. We compare our results with other methods shown in Table III where we also give the recognition rate (87.33%) under a small dictionary size $|\Phi_i^*| = 250$. We can see that our method shows significant improvements in accuracy ($> 10\%$). Even thought there is a small dictionary size (such as $|\Phi_i^*| = 250$), our method is still comparable to other methods. Because the

TABLE III
COMPARISON ON THE UCF SPORTS ACTION DATASET

Papers	Methods	Dictionary Size	Average Accuracy
[30]	Maximum Average Correlation Height	—	69.2%
[42]	Independent subspace analysis	—	86.5%
[1]	class-specific dictionary +cuboid	256	83.8%
[43]	hierarchy of discriminative shape and motion features	300	87.27%
[44]	hough transform-based voting	—	86.6%
[3]	*ME*+STIP	325	81.33%
[6]	*Liu-Shah*+STIP	250	84%
[3]	*Qiu-Jiang*+STIP	308	85.33%
our method	*CSMMI*+STIP	469/250	**98.0%**/87.33%

TABLE IV
COMPARISON ON THE UCF YOUTUBE ACTION DATASET

Papers	Methods	Dictionary Size	Average Accuracy
[31]	cuboid+difussion maps	1000	70.4%
[45]	hybrid features	2000	71.2%
[42]	Independent Subspace Analysis	—	75.8%
[3]	ME+STIP	715	71.1%
[6]	*Liu-Shah*+STIP	624	72.7%
[3]	*Qiu-Jiang*+STIP	678	73.3%
our method	*CSMMI*+STIP	721	**78.6%**

average recognition accuracies obtained from shared dictionary methods are much lower than *CSMMI*, their confusion matrixes are not given. From the above discussion, they serve as proof to that *CSMMI* outperforms both shared dictionary methods and some state-of-the-art methods.

F. UCF YouTube Action Dataset

It contains 11 categories: basketball shooting, cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. This challenging dataset represents the difficulties present in real consumer videos, such as moving/cluttered background, variation in object scale, varied viewpoint and illumination. For each category, the videos are grouped into 25 groups with more than four action clips in it.

We initialized $|\Phi_s^*| = |\Phi_i^*| = 800$ using STIP features and the average accuracy of the initial dictionaries are 67.2% for shared dictionary methods and 74.7% for the class-specific dictionary method. Then the final results of shared and class-specific dictionary methods are shown in Table IV where other state-of-the-art methods are also given. We can see that *CSMMI* improves by 3% the best result from any other method mentioned in Table IV.

G. Hollywood2 Action Dataset

This dataset contains a training set (823 video sequences) and a test set (884 sequences) and includes 12 action classes. Training and test sequences come from different movies and the performance is measured by mean average Precision (mAP) over all classes as in [32]. The authors provide clean and noisy versions of the dataset and we use the clean version.

TABLE V
COMPARISON ON THE HOLLYWOOD2 ACTION DATASET

Papers	Methods	Dictionary Size	mAP
[47]	dense+HOG/HOF	4000	47.4%
[48]	dense trajectories	—	58.3%
[42]	independent subspace analysis	—	53.3%
[46]	compensated descriptors +VLAD representation	—	**62.5%**
[3]	ME+STIP	329	41.3%
[6]	*Liu-Shah*+STIP	415	41.9%
[3]	*Qiu-Jiang*+STIP	394	43.2%
our method	*CSMMI*+STIP	437	**62.1%**

TABLE VI
COMPARISON ON THE KECK ACTION DATASET

Papers	Methods	Static setting	Dynamic setting
[26]	prototype trees	95.2%	91.07%
[49]	Product Manifolds	94.4%	92.3%
[3]	*ME*+shape-motion	91.2%	89.3%
[6]	*Liu-Shah*+shape-motion	94.2%	90.7%
[3]	*Qiu-Jiang*+shape-motion	94.9	92.7%
[3]	*Qiu-Jiang**+shape-motion	97%	—
our method	*CSMMI*+shape-motion	**95.1%**	**93.2%**

We first set $|\Phi_s^0| = |\Phi_i^0| = 500$ for the initial dictionary size using STIP features. The mAPs using initial dictionaries is 39.1% for shared dictionary method and 58.7% for the class-specific dictionary method. Then, we learn Φ_i^* by *CSMMI* (using ε) and the corresponding average dictionary size is 437. The final results are shown in Table V where some state-of-the-art methods on these dataset are given. We can see that *CSMMI* outperforms three shared dictionary-based methods (at least 18% improvement) and is comparable to other state-of-the-art methods such as vector of local aggregated descriptors (VLAD) representation [46].

H. Keck Gesture Dataset

Keck gesture dataset consists of 14 different gesture classes which are military signals. The full list of gestures is: turn left, turn right, attention left, attention right, attention both, stop left, stop right, stop both, flap, start, go back, close distance, speed up and come near. Each gesture is performed by there persons who repeat each gesture three times. The training set including 126 video sequences are captured with the simply and static background while the test set including 168 video sequences contains dynamic an cluttered backgrounds.

We follow the experimental protocol proposed by Jiang et al. [26] for both static and dynamic settings. In the static background setting, the experiments are based on leave-one-person-out cross-validation. As for the dynamic environment, the gestures acquired from the static background are used for training while the gestures collected from the dynamic environment are the test videos. The average rates for both static and dynamic backgrounds are reported in Table VI where we used the shape-motion feature [26] and set $|\Phi_s^0| = |\Phi_i^0| = 600$. We can see that the accuracy rate of our method is slightly lower than that of *Qiu-Jiang**[3] for the static setting.

[3] *Qiu-Jiang* and *Qiu-Jiang** are the same method. In our paper, *Qiu-Jiang*, *ME*, *Liu-Shah* results are based on our own implementations while Qiu-Jiang* results are derived from [3].

TABLE VII
COMPARISON WITH STATE-OF-THE-ART METHODS ON ONE-SHOT LEARNING CHALEARN GESTURE DATASET($devel01 \sim devel20$)

Papers	Methods	Dictionary Size	Average LD
[49]	Product Manifolds	—	0.2873
[52]	Temporal Bayesian Model	—	0.2409
[21]	BoF+3D EMoSIFT	2193	0.1945
[3]	*ME*+3D EMoSIFT	164/1273	0.4851/0.241
[6]	*Liu-Shah*+3D EMoSIFT	146/1637	0.4132/0.223
[3]	*Qiu-Jiang*+3D EMoSIFT	171/1455	0.4045/0.2075
our method	*CSMMI*+3D EMoSIFT	182	**0.1876**

That is because we used a small dictionary size ($|\Phi_i^0| = 600$) in *CSMMI* but *Qiu-Jiang** used a large dictionary size ($|\Phi_s^0| = 1200$).

I. One-Shot Learning Chalearn Gesture Dataset

The main challenges for this dataset are: 1) there is only one training example per class; 2) occlusions may occur (for instance hand covering each other or covering the face). We evaluate the proposed method on the development dataset ($devel01 \sim devel20$) which consists of 20 batches (totally 2000 gestures). Each batch is made of 47 gesture videos and splitted into a training set and a test set. Every test video contains 1 to 5 gestures. Detailed descriptions of gesture data can be found in [33]. This dataset provides both RGB and depth video clips (320×240 pixels).

We used 3D EMoSIFT feature which fuses RGB and depth data. That is because 3D EMoSIFT has revealed promising results on human activity and gesture recognition from RGB-D data [21], [50]. Owing to some video clips including multiple gestures, we first adopted dynamic time warping [21], [51] to achieve temporal segmentation. That means we can get isolated gestures from a test video. Then, we extracted 3D EMoSIFT features for each isolated gesture.

We first set $|\Phi_s^0| = |\Phi_i^0| = 200$. Then the class-specific dictionary Φ_i^* is learned via *CSMMI* (using ϵ) and the corresponding average dictionary is 182. The final results are shown in Table VII where we also give other three additional results in published papers [21], [49], and [52]. In addition, we also set $|\Phi_s^0| = 2000$ for shared dictionary methods and the best results with their corresponding dictionary sizes are shown in Table VII where *CSMMI* outperforms shared dictionary methods in terms of accuracy and efficiency.

To show more detailed information, The LD for each batch is illustrated in Fig. 12. We can see that both class-specific and shared dictionary methods do not work well on $devel03$ batch. That is because 3D EMoSIFT only captures discriminative features when gestures are in motion while there are some static gestures (or postures) on $devel03$ batch. Some postures from $devel03$ batch are shown in Fig. 13. Future work will focus on combining both appearance and motion features to improve recognition rates.

J. Multi-Modal Gesture Dataset

Multi-modal gesture dataset [34] has been recently released. There are 20 gesture classes and the training data has

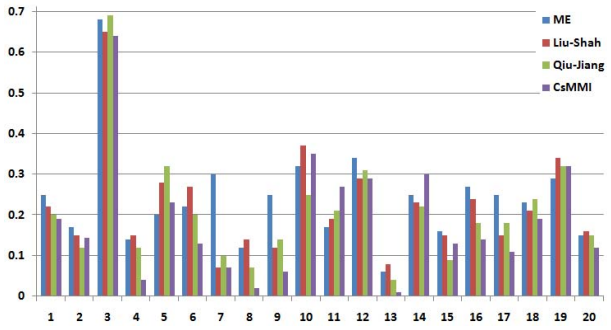


Fig. 12. The performances of each batch ($devel01 \sim devel20$) on one-shot-learning Chalearn dataset. The edit distance and dictionary size of each batch are given: *ME* (0.241,1273); *Liu-Shah* (0.223,1637); *Qiu-Jiang* (0.2075,1455); *CSMMI* (0.1876,182).

Fig. 13. Some static gestures from $devel03$ batch on one-shot learning Chalearn gesture dataset.

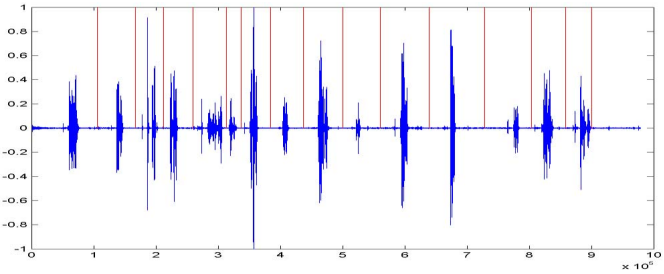


Fig. 14. An illustration of temporal segmentation by the audio file, where the red lines indicate the segment position and blue lines indicate audio data.

393 segments of continuous video, which contain 7754 Italian gestures (Italian cultural/anthropological signs were considered). Each section includes an RGB video, a depth video, an audio file and skeleton information of every frame. The validation data has 287 sessions, which correspond to 3362 Italian gestures.

In this dataset, gestures are continuously performed. It's hard to achieve temporal segmentation via RGB-D data. Here, we use the audio data for this purpose. It is supposed that the audio will return to silence before saying a new gesture by an actor. Therefore, we can detect the silence part [53] in audio file to achieve temporal segmentation. Fig. 14 illustrates the temporal segmentation, where the red lines indicate the detected positions pos_audio from audio files. Because the sample rates between audio files and videos are different, pos_audio are not true positions in videos. The true positions pos_video in videos can be obtained as, $pos\_video = pos\_audio/step$, where $step = fs\_audio/fs\_video$, fs_audio and fs_video represent the sample rate of audio and video files, respectively. In the dataset, $fs\_audio = 16000$ Hz, $fs\_video = 20$ Hz.

Fig. 15. The three samples are represented the same gesture named 'vattene' in Italian. But it shows very different appearance. Any hand of the actor (left hand, right hand or both hands) can perform 'vattene'.

TABLE VIII
COMPARISON ON MULTI-MODAL GESTURE DATASET

		"true segment"		"predicted segment"	
Papers	Methods	Dictionary Size	Average LD	Dictionary Size	Average LD
[3]	*ME* +HOJ3D	695	0.4991	589	0.8537
[6]	*Liu-Shah* +HOJ3D	721	0.4731	694	0.7931
[3]	*Qiu-Jiang* +HOJ3D	697	0.4261	715	0.8120
our method	*CSMMI* +HOJ3D	782	**0.1913**	693	**0.5583**

We first extracted 3D EMoSIFT features from RGB-D data. But it did not yield good results with *CSMMI*. That is because one gesture class can be presented by different hands of an actor as shown in Fig. 15 while 3D EMoSIFT feature is sensitive to the direction of motion [21], [54]. Therefore, we use HOJ3D feature which is view invariant [25].

Because this dataset includes noise gestures (gestures do not contained in the vocabulary), we used training data to learn 21 initialized dictionaries ($|\Phi_i^0| = 1000$) (The noise gestures are treated as a new class). Then we test the dictionary-based methods on validation data. We experimented with two settings. In the "true segment" setting, we predicted the labels using true temporal segmentation. In the "predicted segment" setting, we first used audio files to predict temporal segmentation and then applied shared-dictionary methods or *CSMMI* to achieve gesture recognition. The results and comparisons can be found in Table VIII. We can see that our method can get the best performances in both "true segment" and "predicted segment" settings. It is not surprising that the model performs better in the "true segment" setting, because it has used true temporal segmentation.

K. Discussion

In this section, we critically analyze the results obtained for the proposed method on both synthetic data and public datasets. From the experimental results, the proposed method compares favorably to shared dictionary methods and other state-of-the-art approaches. However, there are still some interesting observations.

1) Running Time: We examined the computational cost of *CSMMI* and compared it with other dictionary-based methods, including *ME*, *Liu-Shah* and *Qiu-Jiang*. Our hardware configuration is 3.30-GHz CPU and 8-GB RAM. Table IX shows that the CPU times spent on the training and testing (recognition) phases using MATLAB on KTH action dataset.

TABLE IX
RUNNING TIMES (IN SECONDS) ON KTH ACTION DATASET

Method	*ME*	*Liu-Shah*	*Qiu-Jiang*	*CsMMI*
Training	44	47	45	1733
Recognition	0.031	0.029	0.027	0.031

Note that five samples of each class were selected for training, and $|\Phi_i^0| = |\Phi_s^0| = 200$. And all selected dictionary items are of size 140 ($|\Phi_i^*| = |\Phi_s^*| = 140$).

The computational complexity of the proposed method for training is generally larger than that of other dictionary-based methods as it has to learn one dictionary for each class. In practical applications, training is usually an offline process. Hence, the recognition time is usually more of our concern than the training time. As shown in Table IX,[4] it allows us to reduce the recognition time of the proposed method and make it comparable to that of other approaches. Moreover, the proposed algorithm can be made more computationally efficient than other approaches, including *ME*, *Liu-Shah* and *Qiu-Jiang*, because it requires a smaller number of dictionary size than other approaches to attain similar or higher performance (see Table VII).

2) Choosing Initial Dictionary Size: The theory of sparse coding and dictionary learning are in a developing stage and the problems to select a initial dictionary size are still open issues [1], [21]. For shared dictionary usually used in BoF model, the authors [47] suggested that the dictionary size is usually set to 4000, which has shown to empirically give good results for a wide range of datasets. In our experiments, we also determine the initial dictionary size for *CSMMI* based on empirical results. When we use STIP features, the initial dictionary sizes for the five action datasets are range from 500 to 800. From our experimental results, when the initial dictionary size for *CSMMI* is not larger than 1000 in most cases, it usually reveals good results.

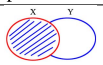
VI. CONCLUSION

In this paper, we present a class-specific dictionary learning approach via information theory for action and gesture recognition. First, the aim of *CSMMI* is to select dictionary items that are highly correlated to a specific class and less correlated to other classes. Second, in the initial class-specific dictionary learning stage, each action class is modeled independently of the others and hence the painful repetition of the training process when a new class is added is no longer necessary. This also indicates the possibility of parallel implementation to speed up the processing time in the recognition stage. Besides, we propose *submodularity* to reduce the complexity of *CSMMI*. Extensive experiments demonstrate that the proposed algorithm has obtained impressive performances.

Although the proposed method has achieved promising results, there are several avenues, which could be explored in the future, for futher improvement, including: how to initialize dictionary items candidates more efficiently and whether there

[4]Because we are using one dictionary per each class, we were able to use the MATLAB parallel computing toolbox in the recognition stage of *CSMMI*.

TABLE X
THE RELATIONSHIPS OF THE ENTROPY, JOINT ENTROPY, CONDITIONAL ENTROPY AND MUTUAL INFORMATION

Name	Symbol	Relationship	graphic representation
Entropy	$H(X)$	$H(X) = H(X\|Y) + I(X;Y)$	
Conditional entropy	$H(X\|Y)$	$H(X\|Y) = H(X \cup Y) - H(Y)$	
Joint entropy	$H(X \cup Y)$	$H(X \cup Y) = H(X) + H(Y\|X)$	
Mutual information	$I(X;Y)$	$I(X;Y) = H(X) - H(X\|Y)$	

is a criterion to learn dictionary with different sizes for different classes in the initial dictionary learning stage. Futher work also includes exploring the optimization of various parameters of the method, such as sparsity for sparse coding. We also intend to apply the proposed method to other classification tasks such as face or object recognition in our future works.

APPENDIX A

We give some concepts about entropy from information theory. The formula derivations in the manuscript will use these basic concepts, which are given in Appendix B.

The entropy $H(X)$ measures the uncertainty inherent in the distribution of a random variable X. Joint entropy $H(X,Y)$ and conditional entropy $H(X|Y)$ are simple extensions that measure the uncertainty in the joint distribution of a pair of random variables X, Y, and the uncertainty in the conditional distribution of a pair of random variables X, Y, respectively. The mutual information $I(X;Y)$ measures how much the realization of random variable Y tells us about the realization of X. More detailed information can be found in [27]. The relationships are shown in Table X.

APPENDIX B

In this appendix, we first give a proposition and its proof which will be used to simplify inter-class MI term τ_2.

Proposition 1: For three variables X, Y, Z, $H(X \cup Z|Y) - H(X|Y) = H(Z|(X \cup Y))$.

Proof: To more easily understand for readers, we also give some graphic representations in Fig. 16. The formula derivations are followed,

$$\begin{aligned} &H((X \cup Z)|Y) - H(X|Y) \\ &\quad = [H(X \cup Z \cup Y) - H(Y)] - [H(X \cup Y) - H(Y)] \\ &\quad = H(X \cup Z \cup Y) - H(X \cup Y) \\ &\quad = H(Z \cup (X \cup Y)) - H(X \cup Y) \\ &\quad = H(Z|(X \cup Y)) \end{aligned}$$

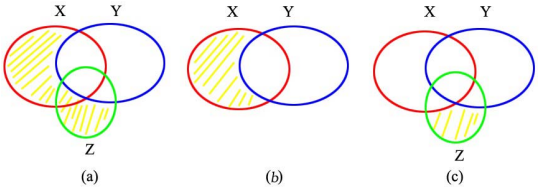


Fig. 16. The graphic representations (the yellow lines denote the object region): (a) $H((X \cup Z)|Y)$; (b) $H(X|Y)$; (c) $H(Z|(X \cup Y))$.

Then we simplify Eq. 4 using formulas of Table X as follow:

$$\begin{aligned} \tau_1 &= H(\Phi_i^* \cup \phi_i) - H((\Phi_i^* \cup \phi_i)|\overline{\Phi_i^*}) \\ &\quad -[H(\Phi_i^*) - H(\Phi_i^*|(\overline{\Phi_i^*} \cup \phi_i))] \\ &= H(\Phi_i^* \cup \phi_i) - [H(\Phi_i^0) - H(\overline{\Phi_i^*})] \\ &\quad -[H(\Phi_i^*) - H(\Phi_i^0) + H(\overline{\Phi_i^*} \cup \phi_i)] \\ &= [H(\Phi_i^* \cup \phi_i) - H(\Phi_i^*)] - [H(\overline{\Phi_i^*} \cup \phi_i) - H(\overline{\Phi_i^*})] \\ &= H(\phi_i|\Phi_i^*) - H(\phi_i|\overline{\Phi_i^*}) \end{aligned} \tag{15}$$

where $\Phi_i^0 = \{\Phi_i^*, \overline{\Phi_i^*}, \phi_i\}$, i is the index of the i^{th} class.

$$\begin{aligned} \tau_2 &= H(\Phi_i^* \cup \phi_i) - H((\Phi_i^* \cup \phi_i)|\Phi_l) \\ &\quad -[H(\Phi_i^*) - H(\Phi_i^*|\Phi_l)] \\ &= H(\Phi_i^*) + H(\phi_i|\Phi_i^*) - H((\Phi_i^* \cup \phi_i)|\Phi_l) \\ &\quad -[H(\Phi_i^*) - H(\Phi_i^*|\Phi_l)] \\ &= H(\phi_i|\Phi_i^*) - [H((\Phi_i^* \cup \phi_i)|\Phi_l) - H(\Phi_i^*|\Phi_l)] \\ &= H(\phi_i|\Phi_i^*) - H(\phi_i|(\Phi_l \cup \Phi_i^*)) \end{aligned} \tag{16}$$

In the last two rows of Eq. 16, we can simplify $H((\Phi_i^* \cup \phi_i)|\Phi_l) - H(\Phi_i^*|\Phi_l) = H(\phi_i|(\Phi_l \cup \Phi_i^*)$ using Proposition 1. Then, we can obtain the objective function of Eq. 5: $\arg\max_{\phi_i \in \Phi_i^0 \setminus \Phi_i^*} H(\phi_i|(\Phi_l \cup \Phi_i^*)) - H(\phi_i|\overline{\Phi_i^*})$, where

$$\begin{aligned} &H(\phi_i|(\Phi_l \cup \Phi_i^*)) - H(\phi_i|\overline{\Phi_i^*}) \\ &\quad = \frac{1}{2}\log(2\pi e\sigma^2_{\phi_i|(\Phi_l \cup \Phi_i^*)}) - \frac{1}{2}\log(2\pi e\sigma^2_{\phi_i|\overline{\Phi_i^*}}) \\ &\quad = \frac{1}{2}\log(\frac{\sigma^2_{\phi_i|(\Phi_l \cup \Phi_i^*)}}{\sigma^2_{\phi_i|\overline{\Phi_i^*}}}) \\ &\quad = \frac{1}{2}\log(\frac{\kappa_{(\phi_i,\phi_i)} - \kappa_{(\phi_i,\Phi_l \cup \Phi_i^*)}\kappa^{-1}_{(\Phi_l \cup \Phi_i^*,\Phi_l \cup \Phi_i^*)}\kappa_{(\Phi_l \cup \Phi_i^*,\phi_i)}}{\kappa_{(\phi_i,\phi_i)} - \kappa_{(\phi_i,\overline{\Phi_i^*})}\kappa^{-1}_{(\overline{\Phi_i^*},\overline{\Phi_i^*})}\kappa_{(\overline{\Phi_i^*},\phi_i)}}) \end{aligned} \tag{17}$$

Therefore, the final objective function is,

$$\arg\max_{\phi_i \in \Phi_i^0 \setminus \Phi_i^*} \frac{\kappa_{(\phi_i,\phi_i)} - \kappa_{(\phi_i,\Phi')}\kappa^{-1}_{(\Phi',\Phi')}\kappa_{(\Phi',\phi_i)}}{\kappa_{(\phi_i,\phi_i)} - \kappa_{(\phi_i,\overline{\Phi_i^*})}\kappa^{-1}_{(\overline{\Phi_i^*},\overline{\Phi_i^*})}\kappa_{(\overline{\Phi_i^*},\phi_i)}}$$

where $\Phi' = \Phi_l \cup \Phi_i^*$.

ACKNOWLEDGMENT

The authors would like to thank Qiang Qiu, University of Maryland, who readily provided the code for three shared-dictionary-based methods (*ME*, *Liu-Shah* and *Qiu-Jiang*).

REFERENCES

[1] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1576–1588, Aug. 2012.

[2] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[3] Q. Qiu, Z. Jiang, and R. Chellappa, "Sparse dictionary-based representation and recognition of action attributes," in *Proc. IEEE ICCV*, Nov. 2011, pp. 707–714.

[4] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 7, pp. 1294–1309, Jul. 2009.

[5] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[6] J. Liu and M. Shah, "Learning human actions via information maximization," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.

[7] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, p. 16, 2011.

[8] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[9] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image sesarch," in *Proc. IEEE ICCV*, vol. 2. Oct. 2005, pp. 1816–1823.

[10] H. Jégou, M. Douze, and C. Schmid, "Packing bag-of-features," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2357–2364.

[11] H. Su, M. Sun, L. Fei-Fei, and S. Savarese, "Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, Sep/Oct. 2009, pp. 213–220.

[12] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2007, pp. 1–8.

[13] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Proc. IEEE ICCV*, Nov. 2011, pp. 1331–1338.

[14] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2011, pp. 1697–1704.

[15] W. Dong, X. Li, D. Zhang, and G. Shi, "Sparsity-based image denoising via dictionary learning and structural clustering," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2011, pp. 457–464.

[16] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2008, pp. 1–8.

[17] J. Zheng and Z. Jiang, "Learning view-invariant sparse representations for cross-view action recognition," in *Proc. IEEE ICCV*, Dec. 2013, pp. 533–547.

[18] L. Yang, R. Jin, R. Sukthankar, and F. Jurie, "Unifying discriminative visual codebook generation with classifier training for object category recognition," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2008, pp. 1–8.

[19] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. 2nd Joint IEEE Int. Workshop Vis. Surveill. Perform. Eval. Tracking Surveill.*, Oct. 2005, pp. 65–72.

[20] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE ICCV*, Nov. 2011, pp. 543–550.

[21] J. Wan, Q. Ruan, S. Deng, and W. Li, "One-shot learning gesture recognition from RGB-D data using bag of features," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 2549–2582, 2013.

[22] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.*, vol. 79, no. 3, pp. 299–318, 2008.

[23] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies," *J. Mach. Learn. Res.*, vol. 9, pp. 235–284, Jun. 2008.

[24] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, 2005.

[25] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Conf. CVPRW*, Jun. 2012, pp. 20–27.

[26] Z. Jiang, Z. Lin, and L. S. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 533–547, Mar. 2012.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[28] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. 10th IEEE ICCV*, vol. 2. Oct. 2005, pp. 1395–1402.

[29] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.

[30] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.

[31] J. Liu, Y. Yang, I. Saleemi, and M. Shah, "Learning semantic features for action recognition via diffusion maps," *Comput. Vis. Image Understand.*, vol. 116, no. 3, pp. 361–377, 2012.

[32] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 2929–2936.

[33] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, and H. J. Escalante, "Chalearn gesture challenge: Design and first results," in *Proc. IEEE Int. Conf. CVPRW*, Jun. 2012, pp. 1–6.

[34] S. Escalera *et al.*, "Chalearn multi-modal gesture recognition 2013: Grand challenge and workshop summary," in *Proc. ACM Int. Conf. Multimodal Interact.*, Dec. 2013, pp. 365–368.

[35] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.

[36] Z. Jiang, G. Zhang, and L. S. Davis, "Submodular dictionary learning for sparse coding," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 3418–3425.

[37] K. Guo, P. Ishwar, and J. Konrad, "Action recognition from video using feature covariance matrices," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2479–2494, Jun. 2013.

[38] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.

[39] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 288–303, Feb. 2010.

[40] I. N. Junejo, E. Dexter, I. Laptev, and P. Pérez, "View-independent action recognition from temporal self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 172–185, Jan. 2011.

[41] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal localization and categorization of human actions in unsegmented image sequences," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 1126–1140, Apr. 2011.

[42] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 3361–3368.

[43] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 2046–2053.

[44] A. Yao, J. Gall, and L. Van Gool, "A hough transform-based voting framework for action recognition," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 2061–2068.

[45] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild'," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1996–2003.

[46] M. Jain, H. Jégou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 2555–2562.

[47] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2009.

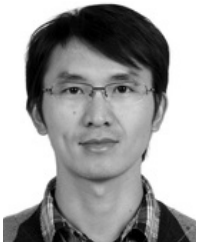
[48] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 3169–3176.

[49] Y. M. Lui, "Human gesture recognition on product manifolds," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 3297–3321, 2012.

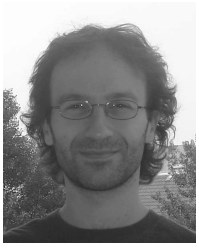
[50] Y. Ming, Q. Ruan, and A. Hauptmann, "Activity recognition from RGB-D camera with 3D local spatio-temporal features," in *Proc. IEEE ICME*, Jul. 2012, pp. 344–349.

[51] I. Guyon, V. Athitsos, P. Jangyodsuk, H. Escalante, and B. Hamner, "Results and analysis of the chalearn gesture challenge 2012," in *Proc. Int. Workshop Adv. Depth Image Anal. Appl.*, vol. 7854. 2013, pp. 186–204.

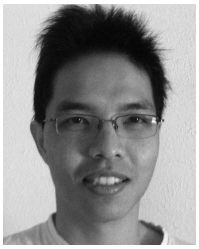
[52] M. R. Malgireddy, I. Inwogu, and V. Govindaraju, "A temporal Bayesian model for classifying, detecting and localizing activities in video sequences," in *Proc. IEEE Int. Conf. CVPRW*, Jun. 2012, pp. 43–48.
[53] T. Giannakopoulos. (2010). *A Method for Silence Removal and Segmentation of Speech Signals, Implemented in Matlab* [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/28826-silence-removal-in-speech-signals
[54] J. Wan, Q. Ruan, W. Li, G. An, and R. Zhao, "3D SMoSIFT: Three-dimensional sparse motion scale invariant feature transform for activity recognition from RGB-D videos," *J. Electron. Imag.*, vol. 23, no. 2, pp. 023017-1–023017-14, Apr. 2014.



Jun Wan received the B.S. degree from the China University of Geosciences, Beijing, China, in 2008, the master's degree from the School of Electronic and Information Engineering Department, Beijing Jiaotong University, Beijing, in 2009, where he is currently pursuing the Ph.D. degree with the Institute of Information Science. He was a recipient of the the 2012 ChaLearn One-Shot-Learning Gesture Challenge Award, sponsored by Microsoft, at ICPR 2012. He was also a recipient of the 2013 Best Paper Award from the Institute of Information Science. His research interests include computer vision, machine learning, in particular, gesture and action recognition, hand tracking, and segmentation.



Vassilis Athitsos received the B.S. degree in mathematics and the M.S. degree in computer science from the University of Chicago, Chicago, IL, USA, in 1995I and 1997, respectively, and the Ph.D. degree in computer science from Boston University, Boston, MA, USA, in 2006. From 2005 to 2006, he was a Researcher with Siemens Corporate Research, Princeton NJ, USA, where he was developing methods for database-guided medical image analysis. From 2006 to 2007, he was a Post-Doctoral Research Associate with the Department of Computer Science, Boston University. Since 2007, he has been a Faculty Member with the Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA, where he currently serves as an Associate Professor. His research interests include computer vision, machine learning, and data mining. His recent work has focused on gesture and sign language recognition, detection and tracking of humans using computer vision, efficient similarity-based retrieval in multimedia databases, and shape modeling and detection. His research has been supported by the National Science Foundation, including an NSF CAREER Award.



Pat Jangyodsuk is currently purusing the Ph.D. degree with the University of Texas at Arlington, Arlington, TX, USA. He received the bachelor's degree in computer engineering from Kasetsart university, Bangkok, Thailand, and the master's degree in computer science from San Jose State University, San Jose, CA, USA. His research interest mainly focuses on gesture recognition problem, or to be more specific American sign language recognition

Hugo Jair Escalante received the Ph.D. degree in computer science from the Instituto Nacional de Astrofisica, Optica y Electronica, Puebla, Mexico, where he is currently an Associate Researcher. He is a member of the Mexican System of Researchers since 2011, and the Director of ChaLearn, the Challenges in Machine Learning Organization from 2011 to 2014. His main research interests are on machine learning and computational intelligence with applications in text mining and high-level computer vision.

Qiuqi Ruan received the B.S. and M.S. degrees from Northern Jiaotong University, Beijing, China, in 1969 and 1981, respectively. From 1987 to 1990, he was a Visiting Scholar with the University of Pittsburgh, Pittsburgh, PA, USA, and the University of Cincinnati, Cincinnati, OH, USA. He is currently a Professor and a Doctorate Supervisor with the Institute of Information Science, Beijing Jiaotong University, Beijing. He is the IEEE Beijing Section Chairman. He has authored and co-authored eight books and more than 350 technical papers in the image processing and information science, and holds one invention patent. His research interests include digital signal processing, computer vision, pattern recognition, and virtual reality.

Isabelle Guyon is an Independent Consultant, and specialized in statistical data analysis, pattern recognition, and machine learning. Her areas of expertise include computer vision and bioinformatics. Her recent interest is in the applications of machine learning to the discovery of causal relationships. Prior to starting her consulting practice in 1996, she was a Researcher with AT&T Bell Laboratories, Murray Hill, NJ, USA, where she pioneered applications of neural networks to pen computer interfaces, and co-invented support vector machines (SVM), a machine learning technique, which has become a textbook method. She is also the primary inventor of SVM-RFE, a variable selection technique based on SVM. The SVM-RFE paper has thousands of citations and is often used as a reference method against which new feature selection methods are benchmarked. She also authored a seminal paper on feature selection that received thousands of citations. She organized many challenges in machine learning over the past few years supported by the EU network Pascal2, NSF, and DARPA, with prizes sponsored by Microsoft, Google, and Texas Instrument. She received the Ph.D. degree in physical sciences from the University Pierre and Marie Curie, Paris, France. She is the President of ChaLearn, a nonprofit dedicated to organizing challenges, the Vice President of the Unipen foundation, an Adjunct Professor with New York University, New York, NY, USA, an Action Editor of the *Journal of Machine Learning Research*, and an Editor of the *Challenges in Machine Learning* book series of Microtome.