Scale Adaptive Dictionary Learning

Cewu Lu, *Member, IEEE*, Jianping Shi, *Student Member, IEEE*, and Jiaya Jia, *Senior Member, IEEE*

***Abstract*— Dictionary learning has been widely used in many image processing tasks. In most of these methods, the number of basis vectors is either set by experience or coarsely evaluated empirically. In this paper, we propose a new scale adaptive dictionary learning framework, which jointly estimates suitable scales and corresponding atoms in an adaptive fashion according to the training data, without the need of prior information. We design an atom counting function and develop a reliable numerical scheme to solve the challenging optimization problem. Extensive experiments on texture and video data sets demonstrate quantitatively and visually that our method can estimate the scale, without damaging the sparse reconstruction ability.**

***Index Terms*— Dictionary learning, sparse coding, sparse representation, image restoration.**

I. Introduction

SPARSE dictionary learning [1] aims to construct dictionaries according to specific input visual data. It gives rise to sparse representation of images patches or video volumes using only a few atoms and has become very popular in these years as it can be employed in solving many image processing problems [2]–[7].

A dictionary contains many atoms in general. Its scale is highly variable, ranging from hundreds to hundreds of thousands in different applications. Experienced developers need a few tryouts or fix it to a number s/he feels comfortable with. For example, in [1], [5], and [8], the scale is set according to experience. In [9], three different dictionary scales are tested.

In terms of scale determination, previous approaches are either time consuming or requiring extensive knowledge. It is especially inconvenient when dealing with applications that involve processing large-scale data or learning many dictionaries at the same time.

For example, in texture synthesis illustrated in Fig. 1, texture data have different dictionary scales, which depend on how informative structures are. For the simple brick texture, 23 dictionary atoms are enough to describe structure variation. On the contrary, for the "crowd" image, its complex patterns lead to a dictionary with 189 atoms. These numbers are not intuitive for humans to be aware of. If the dictionary scale can be determined automatically during optimization, visual data can be processed effectively without needing extensive human experience or prior knowledge.

Manuscript received March 19, 2013; revised August 17, 2013; accepted October 15, 2013. Date of publication October 28, 2013; date of current version January 9, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Richard J. Radke.

The authors are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: lucewu06@gmail.com; jpshi@cse.cuhk.edu.hk; leojia@cse.cuhk.edu.hk).

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.2013.2287602

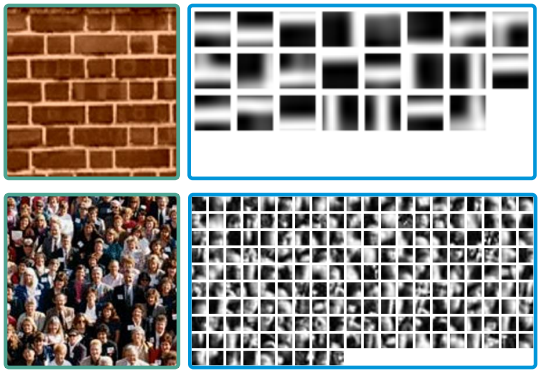


Fig. 1. Two images and their suitable dictionaries used in example-based inpainting and image completion.

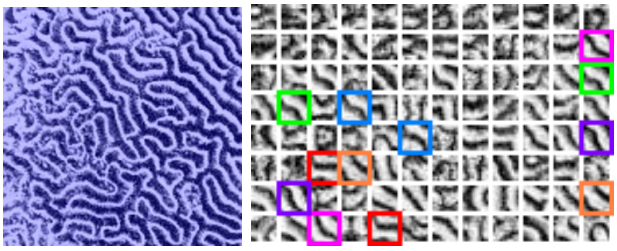


Fig. 2. A texture image and its patch dictionary constructed in [2]. The patches whose correlation is larger than 0.99 are marked with the same color.

It has also been found if the scale of a dictionary deviates much from what it should be, the resulting atoms may not be sufficiently informative or contain many similar or even repeated atoms. The latter case could slow down the testing procedure. An example in Fig. 2 demonstrates that the method of [2] could suffer from redundant atoms.

Bayesian sparse models [10] were developed aiming to learn dictionaries in a non-parametric way. Inferring dictionary scales is also achievable. But, as pointed out in [11], these methods may not know whether the Bayesian model is appropriate or not for the data at hand. Further, they generally take heavy computational costs. Ramirez *et al.* [11] employed the Minimum Description Length (MDL) principle to estimate dictionary size using an enumeration scheme. It estimates all possible dictionary scales from one to the maximum value allowed. When the latent dictionary scale is large, this enumeration scheme is not that efficient. Moreover, both Bayesian sparse [10] and MDL [11] models cannot avoid identical and very similar atoms theoretically.

In this paper, we propose a Scale Adaptive Dictionary Learning (SADL) method. Unlike enumeration in MDL [11], it is a unified framework to learn the sparse dictionary rep-

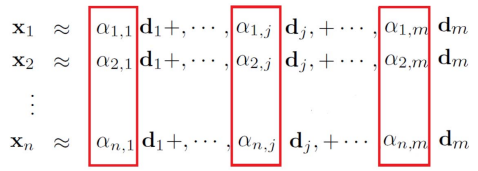


Fig. 3. In sparse linear combination of basis vectors, coefficients α can measure whether one basis vector is used or not. An Atom Indicator Vector (AIV) $\widehat{\boldsymbol{\alpha}}_j$ contains all coefficients in a red rectangle corresponding to basis $\mathbf{d}_j$.

resentation and determine the appropriate number of atoms simultaneously, which has a dissimilarity lower bound for any two atoms theoretically.

Our main contribution is threefold. First, we enable the learnt dictionary scale automatically adaptive to the input data by introducing Atom Indicator Vectors (AIVs) to describe the compactness of output atoms. Second, we prove that our model can lead to a compact dictionary with a nonzero atom-wise distance lower bound. Third, we utilize the Multivariate Moreau Proximal Indicator (MMPI) penalty to solve for SADL efficiently. Our extensive experiments in different visual data manifest that our learnt dictionaries preserve good reconstruction ability and their scales are appropriate.

II. Analysis and Formulation

In this paper, matrices, vectors and sets are in bold capital, bold lower-cased and calligraphic fonts respectively. For a dictionary matrix $\mathbf{D}$, $\mathbf{d}_i$ denotes the i-th column (i.e., an atom) and $\hat{\mathbf{d}}_j$ denotes the j-th row.

Typical dictionary learning is formulated as

$$\min_{\mathbf{D}\in\mathcal{D},\mathcal{A}} \quad \mathbf{E}(\mathbf{D},\mathcal{A}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{2}\|\mathbf{D}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1\right\}, \quad (1)$$

where λ is a regularization parameter. $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is the training data set with sparse coefficients $\mathcal{A} = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n\}$ over dictionary $\mathbf{D}$ in $\mathbb{R}^{m\times k}$. $\|\mathbf{D}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2$ is the data fitting term whereas $\|\boldsymbol{\alpha}_i\|_1$ is sparsity regularization. The problem can be solved via alternatively solving for $\mathbf{D}$ and $\mathcal{A}$. Dictionary $\mathbf{D}$ is restricted to a closed convex set $\mathcal{D}$ following the setting in [9]:

$$\mathcal{D} \triangleq \{\mathbf{D} \in \mathbb{R}^{m\times k} \quad s.t. \quad \forall\, j = 1, \ldots, k, \quad \mathbf{d}_j^T\mathbf{d}_j \le 1\}. \quad (2)$$

In Eq. (1) the size of the dictionary is a free parameter. Originated in the well known model selection criteria, such as AIC and BIC [12], we introduce a model scale penalty. In signal processing, scale penalty is described by a row-sparse norm [13]. We alternatively make use of the corresponding response from $\mathcal{A}$ and define Atom Indicator Vectors (AIVs) $\widehat{\boldsymbol{\alpha}}_j = [\alpha_{1,j}, \ldots, \alpha_{n,j}]$ $(1 \le j \le k)$, where $\alpha_{i,j}$ is the j^{th} element of $\boldsymbol{\alpha}_i$. It is illustrated in Fig. 3. AIVs can measure the importance of individual basis with the count of zero elements. To utilize this clue, we update the original dictionary learning framework to

$$\min_{\mathbf{D}\in\mathcal{D},\mathcal{A},} \quad \mathbf{E}(\mathbf{D},\mathcal{A}) + \mu\sum_{j=1}^{k}\mathbf{I}(\widehat{\boldsymbol{\alpha}}_j), \quad (3)$$

where $\sum_{j=0}^{k}\mathbf{I}(\widehat{\boldsymbol{\alpha}}_j)$ imposes dictionary scale penalty and μ is a balance parameter. The indicator function is defined as

$$\mathbf{I}(\widehat{\boldsymbol{\alpha}}_j) = \begin{cases} 0 & \text{if } \widehat{\boldsymbol{\alpha}}_j = \mathbf{0}_n \\ 1 & \text{otherwise} \end{cases}$$

It outputs 1 for non-zero vectors. Hence, the sum of the indicator functions for all AIV elements, expressed as $\sum_{j=1}^{k}\mathbf{I}(\widehat{\boldsymbol{\alpha}}_j)$, can represent the number of the atoms that are indeed used.

Note that we do not assume that a dictionary originally contains zero vectors. Instead, the objective function within Eq. (3) can automatically control the number of non-zero AIVs in optimization. The linear sparse representation can be adjusted in scale by penalizing $\sum_{j=1}^{k}\mathbf{I}(\widehat{\boldsymbol{\alpha}}_j)$. We name this method as Scale Adaptive Dictionary Learning (SADL). We note that previous model selection methods, such as AIC and BIC [12], aim to compare models. They cannot directly estimate a scale.

A. Dictionary Compactness and Scale Adaptation

High dictionary compactness makes learnt atoms discriminative. There are approaches, such as [14], that add extra discriminative terms to accomplish this goal. But these methods still pre-define the dictionary size, independently from the data at hand. Our framework can ideally capture this compactness property. We prove in what follows that it can avoid identical or very similar atoms in dictionary learning. We also show that the Euclidian distance between any two learnt atoms in our results has a nonzero lower bound. These conditions have never been discussed in this field. They are also not necessarily satisfied in prior models.

Theorem 1: Assuming $\{\mathbf{D}^*, \mathcal{A}^*\}$ is the optimal solution of Eq. (3), any two atoms $\mathbf{d}_v^*$ and $\mathbf{d}_u^*$ with $\mathbf{I}(\widehat{\boldsymbol{\alpha}}_v^*) = 1$ and $\mathbf{I}(\widehat{\boldsymbol{\alpha}}_u^*) = 1$ must satisfy

$$\|\mathbf{d}_v^* - \mathbf{d}_u^*\|_2^2 \ge \frac{n\mu\lambda^2}{\kappa\phi^2}, \quad (4)$$

where $\phi = \sum_{i=1}^{n}\{\frac{1}{2}\|\mathbf{x}_i\|_2^2\}$ and $\kappa = 1 + \frac{\lambda}{\sqrt{n\phi}}$.

The proof is given in the Appendix. Theorem 1 indicates that when $\mu = 0$, Eq. (3) degrades to the traditional dictionary learning model expressed in Eq. (1). Given a non-zero μ, Theorem 1 ensures a dissimilarity lower bound for any two atoms, making the learnt dictionary compact.

III. Multivariate Moreau Proximal Indicator

The objective function in Eq. (3) involves multivariate indicator terms $\mathbf{I}(\widehat{\boldsymbol{\alpha}}_j)$. We introduce a novel Multivariate Moreau Proximal Indicator (MMPI) penalty $\Upsilon(\mathbf{a})$ to avail optimization. The MMPI penalty is defined as

$$\Upsilon_\rho(\mathbf{a}) = \min_{\mathbf{t}\in\mathbb{R}^n}\{\rho\|\mathbf{a} - \mathbf{t}\|_2^2 + \mathbf{I}(\mathbf{t})\}. \quad (5)$$

If ρ is sufficiently large, MMPI approaches the multivariate indicator function $\mathbf{I}(\widehat{\boldsymbol{\alpha}}_j)$. To facilitate description, we plot the penalty of Υ_ρ in 1D and 2D under different ρ in Fig. 4. The peak gets sharper with a larger ρ. When $\rho = 1000$, Υ_ρ is nearly identical to the multivariate indicator function $\mathbf{I}$.

With the MMPI penalty, we can optimize the problem with a suitable model scale. Also, the MMPI penalty is quite

different from the Moreau proximal mapping discussed in [15], which involves only univariate indicator functions whereas our MMPI is a multivariate model. We give in Lemma 1 the MMPI solution. That lemma enables optimization in a very nice form.

Lemma 1: The solution to the MMPI penalty $\Upsilon_\rho(\mathbf{a})$ in Eq. (5) is with the form

$$\Upsilon_\rho(\mathbf{a}) = \begin{cases} \rho\|\mathbf{a}\|_2^2 & \text{if } \|\mathbf{a}\|_2^2 < 1/\rho, \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

Proof: We discuss two situations.

- When $\mathbf{t} = \mathbf{0}_n$, it equals to $\rho\|\mathbf{a}\|_2^2$.
- When $\mathbf{t} \neq \mathbf{0}_n$, the optimum is reached when $\mathbf{t} = \mathbf{a}$. In this case, $\Upsilon_\rho(\mathbf{a}) = \min_{\mathbf{t}}\{\rho\|\mathbf{a}-\mathbf{t}\|_2^2 + 1\} = 1$.

In summary, if $\rho\|\mathbf{a}\|_2^2 < 1$, we get $\Upsilon_\rho(\mathbf{a}) = \rho\|\mathbf{a}\|_2^2$. Otherwise, $\Upsilon_\rho(\mathbf{a}) = 1$. ∎

IV. Optimization

We approximate the multivariate indicator term $\mathbf{I}(\widehat{\boldsymbol{\alpha}}_j)$ by the MMPI penalty $\Upsilon_\rho(\widehat{\boldsymbol{\alpha}}_j)$ in Eq. (3) and obtain function

$$\min_{\mathbf{D}\in\mathcal{D},\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\{\|\mathbf{D}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1\} + \mu\sum_{j=1}^{k}\Upsilon_\rho(\widehat{\boldsymbol{\alpha}}_j). \tag{7}$$

It now can be solved efficiently in a two-step scheme, which optimizes $\mathbf{D}$ and $\mathcal{A}$ alternatively. The two steps are referred to as *dictionary update* and *dictionary selective sparse coding* respectively.

A. Dictionary Update

Given the estimated $\mathcal{A}$ in the previous step, we solve

$$\min_{\mathbf{D}\in\mathcal{D}} \mathbf{L}(\mathbf{D}) = \min_{\mathbf{D}\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{2}\|\mathbf{D}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2\right\}. \tag{8}$$

We resort to the classical first-order projected stochastic gradient descent algorithm [16] to compute $\mathbf{D}$. It updates $\mathbf{D}$ iteratively. In each iteration,

$$\mathbf{D} = \prod_{\mathcal{D}}[\mathbf{D} - \delta_t\nabla_{\mathbf{D}}\mathbf{L}(\mathbf{D})],$$

where δ_t is the gradient operator, and $\prod_{\mathcal{D}}$ represents the projector to refine the dictionary in set $\mathcal{D}$.

B. Dictionary Adaptive Sparse Coding

With the estimated $\mathbf{D}$ in the above step, we minimize $\mathcal{A}$ by solving

$$\min_{\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\{\frac{1}{2}\|\mathbf{D}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1\} + \mu\sum_{j}^{k}\Upsilon_\rho(\widehat{\boldsymbol{\alpha}}_j). \tag{9}$$

According to the definition of $\Upsilon_\rho(\cdot)$ in Eq. (5), we rewrite Eq. (9) as

$$\min_{\mathcal{A}} \frac{1}{n}\Big\{\sum_{i=1}^{n}[\|\mathbf{D}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1] + \mu\min_{\mathbf{T}}\sum_{j=1}^{k}[\rho\|\widehat{\boldsymbol{\alpha}}_j - \widehat{\mathbf{t}}_j\|_2^2 + \mathbf{I}(\widehat{\mathbf{t}}_j)]\Big\}. \tag{10}$$

We provide its equivalent formulation as

$$\min_{\mathcal{A},\mathbf{T}} \frac{1}{n}\Big\{\sum_{i=1}^{n}[\|\mathbf{D}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1] + \mu\sum_{j=1}^{k}[\rho\|\widehat{\boldsymbol{\alpha}}_j - \widehat{\mathbf{t}}_j\|_2^2 + \mathbf{I}(\widehat{\mathbf{t}}_j)]\Big\}. \tag{11}$$

Similar to $\widehat{\boldsymbol{\alpha}}_j$, $\widehat{\mathbf{t}}_j$ is the corresponding row vector in $\mathbf{T}$, and $\mathbf{t}_j$ is the column vector. It holds that $\mathbf{T} = [\widehat{\mathbf{t}}_1, \ldots, \widehat{\mathbf{t}}_k]^T = [\mathbf{t}_1, \ldots, \mathbf{t}_n]$.

Since there are two variables $\mathcal{A}$ and $\mathbf{T}$ in Eq. (11), we decompose the problem into two sub ones, both of which have closed form solutions.

1) Updating $\mathcal{A}$: We ignore the constant terms with respect to $\mathcal{A}$ in Eq. (11). The objective function becomes

$$\begin{aligned}&\min_{\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\{\|\mathbf{D}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i)\|_1\} + \mu\sum_{j=1}^{k}\rho\|\widehat{\boldsymbol{\alpha}}_j - \widehat{\mathbf{t}}_j\|_2^2 \\ &= \min_{\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\{\|\mathbf{D}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1 + n\mu\rho\|\boldsymbol{\alpha}_i - \mathbf{t}_i\|_2^2\}. \end{aligned}\tag{12}$$

Functions for different i are independent. We thus solve each separately as

$$\min_{\boldsymbol{\alpha}_i} \|\mathbf{D}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1 + n\mu\rho\|\boldsymbol{\alpha}_i - \mathbf{t}_i\|_2^2. \tag{13}$$

It is a combination of quadratic term and ℓ^1 sparse term, this formation can be solved by iterative shrinkage and thresholding method [17] efficiently.

2) Updating $\mathbf{T}$: We ignore the constant term with respect to $\mathbf{T}$ in Eq. (11) and solve

$$\min_{\widehat{\mathbf{t}}_1\in\mathbf{T}} \sum_{j=0}^{k}\rho\|\widehat{\boldsymbol{\alpha}}_j - \widehat{\mathbf{t}}_j\|_2^2 + \mathbf{I}(\widehat{\mathbf{t}}_j). \tag{14}$$

It can be decomposed to k independent functions with respect to index j. Without loss of generality, we discuss how the j^{th} problem is solved, which is

$$\min_{\widehat{\mathbf{t}}_j} \rho\|\widehat{\boldsymbol{\alpha}}_j - \widehat{\mathbf{t}}_j\|_2^2 + \mathbf{I}(\widehat{\mathbf{t}}_j). \tag{15}$$

It is a standard MMPI penalty, and can be directly solved via Lemma 1.

C. SADL Framework Summary

In summary, starting with a random $\mathbf{D}$, we apply Algorithm 1. In the inner iteration of $\{\mathcal{A}, \mathbf{T}\}$, when the energy

$$E_\rho(\mathbf{D}, \mathcal{A}, \mathbf{T}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{2}\|\mathbf{D}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1\right\} + \mu\sum_{j=1}^{k}[\rho\|\widehat{\boldsymbol{\alpha}}_j - \widehat{\mathbf{t}}_j\|_2^2 + \mathbf{I}(\widehat{\boldsymbol{\alpha}}_j)] \tag{16}$$

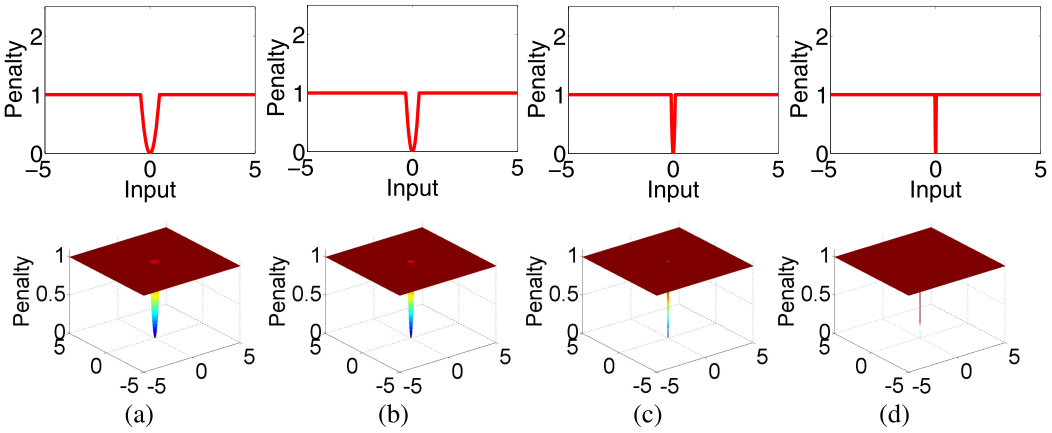


Fig. 4. 1D and 2D plots for Υ_ρ under different ρs. A larger ρ corresponds to a smaller aperture in the plot. (a) $\rho = 5$. (b) $\rho = 10$. (c) $\rho = 100$. (d) $\rho = 1000$.

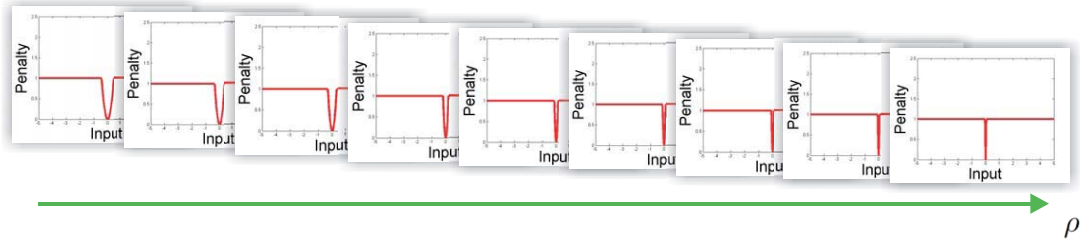


Fig. 5. ρ increases gradually in iterations to make $\Upsilon(\cdot)$ approach $\mathbf{I}(\cdot)$.

Algorithm 1 Scale Adaptive Dictionary Learning (SADL)

input: input data $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$; regularization parameters λ and μ
initialize $\rho = 1$, $t = 1$; generating $\mathbf{D}_0$ randomly.
repeat
  $\mathbf{T}^0 = \mathbf{T}_{t-1}$, $\mathcal{A}^0 = \mathcal{A}_{t-1}$; $i = 0$
  repeat
    with $\mathbf{T}^{i-1}$, solve for $\mathcal{A}^i$ in Eq. (13);
    with $\mathcal{A}^i$, solve for $\mathbf{T}^i$ in Eq. (15);
    $i = i + 1$
  until $E_\rho(\mathbf{D}_{t-1}, \mathbf{T}^i, \mathcal{A}^i)$ converge;
  $\mathbf{T}_t = \mathbf{T}^i, \mathcal{A}_t = \mathcal{A}^i$;
  with $\mathcal{A}_t, \mathbf{D}_{t-1}$, solve for $\mathbf{D}_t$ using gradient descent algorithm [16];
  $\rho \leftarrow 2\rho$, $t = t + 1$;
until $\mathbf{D}_t$ converge or $\rho > 10^5$
$\mathbf{D}^* = \mathbf{D}_t$, $\mathbf{T}^* = \mathbf{T}_t$
return atoms $\{\mathbf{d}_j^* | \mathbf{I}(\widehat{\mathbf{t}}_j^*) = 1\}$ for $\forall j = 1, \ldots, k$.

reaches its limit, the system terminates. The final dictionary consists of atoms $\{\mathbf{d}_j^* | \mathbf{I}(\widehat{\mathbf{t}}_j^*) = 1\}$. The scale is automatically adaptive to input visual data.

D. Analysis and Discussion

1) Convergence Analysis: In Algorithm 1, we increase ρ gradually in iterations as shown in Fig. 5. This scheme, compared to fixing ρ as a large value, warms up the optimization, and has the effect to pull results out of local minima.

According to Eq. (6), the distance between $\mathcal{A}$ and $\mathbf{T}$ reduces in iterations, whose upper-bound in t^{th} iteration is

$$\|\widehat{\boldsymbol{\alpha}}_j - \widehat{\mathbf{t}}_j\|_2^2 \leq \min\left\{\frac{1}{\rho_t}, 0\right\} \leq \frac{1}{2^t}, \tag{17}$$

where ρ_t is the value of ρ in the t^{th} iteration. With improved $\mathcal{A}$ and $\mathbf{T}$, $\mathbf{D}$ is updated until convergence.

2) Parameter Discussion: Parameter λ controls the sparsity. Its value is in $0.2 \sim 0.3$. Its empirical validation on visual data is presented below. μ is the regularization strength. Its effect is to exclude atoms that are least used in the training data. Even if its value is fixed, scale can still be automatically adaptive. Empirically, we set $\mu = 0.002$ in our experiments.

V. Experiments

We conduct extensive experiments to verify our model. In qualitative evaluate, we define "safe dictionary" and "85%-dictionary." *Safe dictionaries* are trained via the traditional method [1], which are with double the number of atoms than those produced in our method. If our dictionaries are similarly effective as these *safe* ones, our learnt dictionary is regarded as complete. Meanwhile, we train dictionaries with 85% of the size determined by our method. We call them 85%-*dictionaries*. If reducing 15% of the atoms significantly increases sparse reconstruction errors, it is obvious that our estimated scale is very close to the lower bound that a dictionary needs to be with.

A. Evaluation on Synthetic Data

This experiment is to manifest that our proposed approximation algorithm solving Eq. (19) can be very similar in terms

Fig. 6. (a)–(g) are the seven texture examples with increasing complexity.

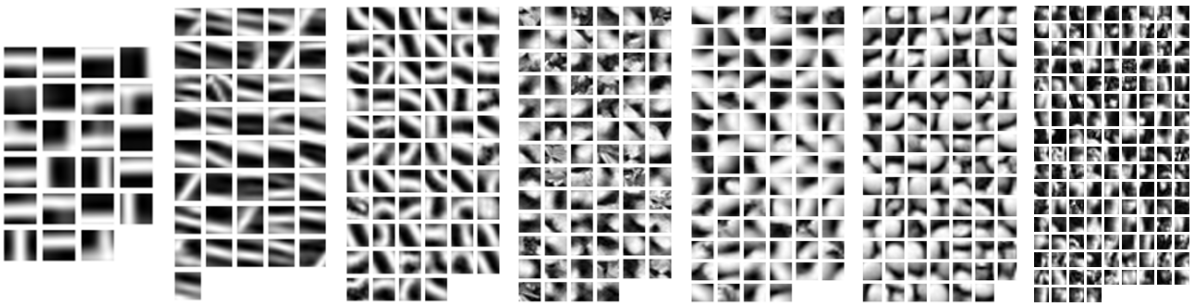


Fig. 7. Dictionaries trained on textures in Fig. 6. Atoms increase from left to right. The numbers are the scales of D. (a) 23. (b) 41. (c) 64. (d) 76. (e) 82. (f) 103. (g) 189.

TABLE I
COMPARISON OF SCALES COMPUTED BY THE TWO MODELS

Ground Truth Scale s	10	50	100	150	200
Ideal model Eq. (18)	11	56	113	159	216
Approximation Eq. (19)	13	62	124	166	235

of performance to the ideal model in Eq. (18).

$$\min_{\mathbf{D}\in\mathcal{D},\mathcal{A},} \frac{1}{n}\sum_{i=1}^{n}\{\|\mathbf{D}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1\} + \mu\sum_{j=1}^{k}\mathbf{I}(\widehat{\boldsymbol{\alpha}}_j). \quad (18)$$

$$\min_{\mathbf{D}\in\mathcal{D},\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\{\|\mathbf{D}\boldsymbol{\alpha}_i - \mathbf{x}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1\} + \mu\sum_{j=1}^{k}\Upsilon_\rho(\widehat{\boldsymbol{\alpha}}_j). \quad (19)$$

We randomly generate a dictionary $\mathbf{D} \in \mathbb{R}^{16\times s}$ and the sparse coefficients $[\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]$ with 20% sparsity ratio, where $n = 40000$ and the dictionary scale s is chosen under different levels from 10 to 200. Our data are generated via $\mathbf{x}_i = \mathbf{D}\boldsymbol{\alpha}_i + \xi_i (1 \le i \le n)$, where ξ_i is the additive Gaussian noise (SNR = 30 dB).

We compare the results produced using the two models under different dictionary scales s. We achieve the solution of Eq. (18) by exhaustively trying all possible scales and finding the best one that gives a small reconstruction error. The results are demonstrated in Table I. It shows that the results from the two models are close enough.

We compare our estimates with those of [11] and [10] with ground truth dictionary scales s ranging from 10 to 1600. The estimation errors are listed in Table III. It shows our estimates are generally more accurate in terms of scales.

Our solver performs favorably with regard to running time. This is because MDL [11] adopting enumeration takes heavy computation as the dictionary scale grows and BDL [10] has to repeatedly solve optimization problems. We list running time of the three methods on a PC (CPU 2.80GHz, RAM 2.96GB) in Table IV. Our method is more efficient due to much less iterations in optimization being performed.

TABLE II
ESTIMATED DICTIONARY SCALES ON THE 25 MTC SETS

MTC set	1	2	3	4	5	6	7
$\lvert\mathbf{D}\rvert$	52	93	135	164	172	182	201
MTC set	8	9	10	11	12	13	14
$\lvert\mathbf{D}\rvert$	232	277	315	346	384	425	462
MTC set	15	16	17	18	19	20	21
$\lvert\mathbf{D}\rvert$	493	522	551	560	578	599	618
MTC set	22	23	24	25			
$\lvert\mathbf{D}\rvert$	639	661	677	702			

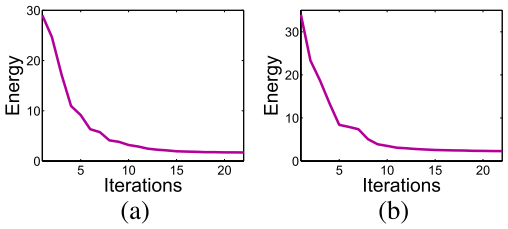


Fig. 8. (a) and (b) demonstrate the energy change during optimization for textures in Fig. 7(a) and (b) respectively. Energy decreases quickly in both examples.

B. Texture Experiments

Sparse representation is very useful in solving many texture-involved problems, such as texture inpainting, synthesis, and classification [18], [19]. We first use them to evaluate our method. In general, structure complexity of texture or the amount of information stored can be coarsely perceived. For

TABLE III

ESTIMATION ERRORS $|s^* - s|/s$ FOR EACH METHOD, WHERE s^* AND s ARE ESTIMATED AND GROUND TRUTH SCALES RESPECTIVELY

Ground Truth Scales s	10	50	100	150	200	400	600	800	1000	1200	1400	1600	Average
MDL [11] scale error (%)	30.0	32.0	35.0	26.6	24.5	22.7	15.5	19.3	19.2	17.5	16.2	23.3	23.5
BDL [10] scale error (%)	20.0	28.0	32.0	23.3	22.5	18.2	19.5	17.7	18.6	19.0	21.3	24.6	22.0
Our SADL scale error (%)	30.0	24.0	24.0	10.6	17.5	10.2	11.6	12.7	12.9	11.9	10.7	13.9	15.8

TABLE IV

RUNNING TIME UNDER DIFFERENT SCALES

Ground Truth Scale s	10	50	100	150	200	400	600	800	1000	1200	1400	1600
MDL [11] (hours)	0.05	0.29	0.56	1.10	1.95	3.75	4.54	5.32	5.20	7.54	8.21	9.38
BDL [10] (hours)	1.37	2.19	3.48	3.33	3.95	4.52	5.05	6.13	6.55	7.24	7.45	8.29
Our SADL (hours)	0.07	0.12	0.21	0.25	0.29	0.42	0.57	0.59	0.63	0.67	0.73	0.78

TABLE V

AVERAGE RECONSTRUCTION ERRORS WITH VARYING SPARSITY IN FIG. 6. "SPARSITY" IS MEASURED AS THE RATIO OF NON-ZERO ATOM NUMBER TO THE TOTAL NUMBER N LEARNT AUTOMATICALLY BY OUR METHOD. "SCALE" DENOTES THE NUMBER OF DICTIONARY ATOMS. WE COMPARE OUR RESULTS TO THOSE WITH DICTIONARY SIZES PRE-DEFINED AS $2N$ AND $0.85N$

Sparsity	Scale	Texture 1	Texture 2	Texture 3	Texture 4	Texture 5	Texture 6	Texture 7
10%	N	0.5388	1.3144	1.7752	1.2959	0.4135	0.7590	1.0292
	$2N$	0.5144	1.2841	1.6594	1.2863	0.3910	0.7422	0.9609
	$0.85N$	0.8568	1.7937	2.3220	1.6851	0.6163	1.0399	1.4658
20%	N	0.4511	1.2949	1.6571	1.2326	0.4056	0.7503	0.9327
	$2N$	0.4346	1.2354	1.5830	1.2100	0.4034	0.7274	0.9035
	$0.85N$	0.7413	1.7917	2.2879	1.6741	0.6060	1.0122	1.3758
40%	N	0.3643	1.2242	1.5512	1.1740	0.4104	0.7060	0.8910
	$2N$	0.3599	1.2077	1.4852	1.1647	0.3940	0.6958	0.8536
	$0.85N$	0.5892	1.7597	2.0988	1.6153	0.5719	0.9744	1.3286

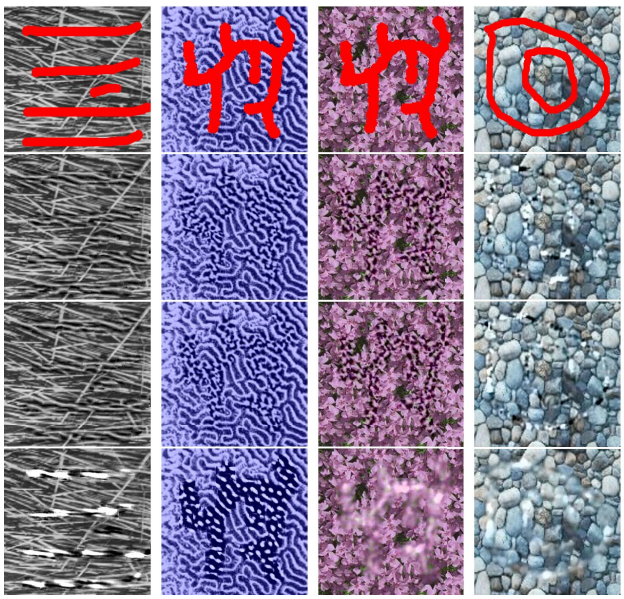


Fig. 9. Inpainting results. First row: damaged images. Second row: inpainting results using our dictionaries. Third row: inpainting results using safe dictionaries. Fourth row: inpainting using 15% smaller dictionaries.

example, in Fig. 6, the left most texture is apparently less complex than the right most ones. So the dictionary size should increase accordingly. In our experiments, we resize texture images to 400×400 pixels. Patches in each image are regularly sampled with size 16×16 in an overlapping manner. We compare the resulting dictionary scales and conduct inpainting to evaluate our method.

TABLE VI

MEAN, VARIANCE, MINIMUM, AND MAXIMUM OF 100 ESTIMATED SCALES PRODUCED WITH DIFFERENT INITIALIZATION FOR EACH TEXTURE EXAMPLE

	(a)	(b)	(c)	(d)	(e)	(f)	(g)
Mean	22.9	41.0	63.8	75.9	82.1	103.0	188.9
Variance	0.19	0.13	0.23	0.13	0.16	0.21	0.20
Minimum	24	42	66	78	83	105	191
Maximum	22	40	63	74	80	101	187

TABLE VII

SCALE ESTIMATES UNDER DIFFERENT λ ON THE SEVEN TEXTURES

λ	(a)	(b)	(c)	(d)	(e)	(f)	(g)
0.1	23	41	64	76	82	103	189
0.2	23	41	64	76	82	103	189
0.3	23	41	64	76	81	103	189
0.4	23	41	63	76	81	102	188

1) Performance With Different Starting Points: We use a random dictionary for initialization. Experiments have been conducted to evaluate how sensitive our algorithm is to different starting points. For each texture in Fig. 6, we randomly generate 100 different initial dictionaries, starting from which we produce our results. Statistics are listed in Table VI.

2) Parameter Setting: Two parameters λ and μ are allowed to vary in our method. We show how results are influenced in Tables VII and VIII. These statistics manifest that our method is not vastly sensitive to these parameters when they are reasonably set and thus can use fixed values in general.

TABLE VIII
SCALE ESTIMATES UNDER DIFFERENT μ ON THE SEVEN TEXTURES

μ	(a)	(b)	(c)	(d)	(e)	(f)	(g)
0.0005	23	41	65	77	83	104	189
0.001	23	41	64	76	82	103	189
0.002	23	41	64	76	82	103	189
0.004	23	41	64	76	82	103	189
0.008	23	41	63	76	81	103	188
0.016	23	40	62	75	81	102	188

TABLE IX
ESTIMATED DICTIONARY SCALES BY BDL AND OUR METHOD ON THE SEVEN TEXTURES

	(a)	(b)	(c)	(d)	(e)	(f)	(g)
SADL	23	41	64	76	82	103	189
BDL	198	211	230	212	241	237	244

3) Scale Adaption Evaluation: We apply our method to a set of texture images in Fig. 6. Our experimental results manifest the intuition that the left- and right-most dictionary sizes vary a lot. For the simple brick texture, 23 basis vectors are enough to describe structure variation, as shown in Fig. 7. For the flower image, the texture has more details. Its dictionary size accordingly increases to 76. Finally for the crowd texture, although its resolution is small, the many details lead to a dictionary with 189 atoms, complying with our visual intuition. For each texture, we have 10,000 training patches; the average training time for each texture is 5.90 minutes.

In quantitative evaluation, we calculate and compare average sparse reconstruction errors $\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}-\mathbf{D}\boldsymbol{\beta}\|_2^2$ for all the patches. They are listed in Table V. They indicate that our dictionaries are complete and their scales are close to the lower bounds that the dictionaries need to be with.

We have also experimented with a larger-scale texture dataset [20], which contains 25 texture classes and 40 samples in each class. We index the texture class from $1 \sim 25$ and define 25 mix-texture class (MTC) sets where the i^{th} MTC set includes *all* textures from class $1 \sim i$ – that is, the i^{th} MTC is a subset of $(i+1)^{th}$ MTC. For each MTC, the number of training patches is fixed to 50000; all texture classes contribute equally to the samples. The estimated dictionary scales by our method are shown in Table II. With the increase of patch structure variety, our dictionary size grows from MTC 1 to 25 steadily.

The convergence is guaranteed in our method. In applying patch-based dictionary learning to texture images, the energy in Eq. (7) decreases quickly within a few iterations, as shown in Fig. 8.

Non-parametric Bayesian dictionary learning (BDL) [10] can also estimate the dictionary scale as a byproduct. It is notable that the dictionary compactness cannot be guaranteed in this method. We compare BDL with our SADL framework on 7 textures on texture reconstruction based on dictionary learning. The code of BDL is provided by the authors. Default parameters are used. Initial atom number is set to 256. The compared result is shown in Table IX. Our estimated dictionary scales are much smaller. We also plot the reconstruction errors for texture in Fig. 10(a). The reconstruction errors on 100 randomly selected patches are comparable. Fig. 10(b) shows a resulting dictionary by BDL. Our method can achieve smaller dictionary scale in comparison to BDL under comparable error reconstruction.

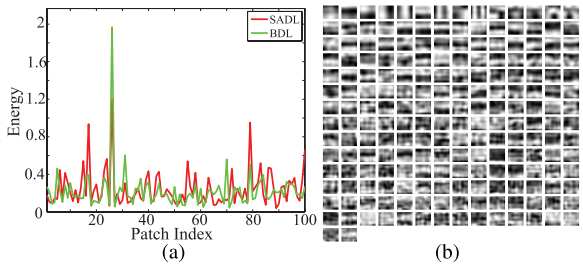


Fig. 10. (a) Sparse reconstruction errors of our dictionaries and those of BDL under the same sparsity degree on the left most brick texture in the seven textures examples. (b) A dictionary learned by BDL for the brick texture.

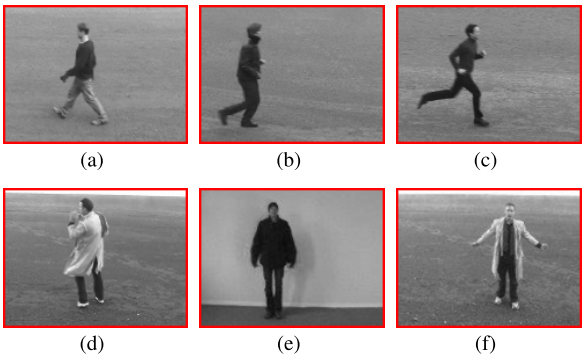


Fig. 11. (a)–(f) are six human action samples in [21].

4) Texture Inpainting: We visually compare the sparse reconstruction results in texture inpainting using our dictionaries and the *safe dictionaries* with double the number of atoms. Reconstruction coefficients of the damaged texture patches are computed via minimizing $\frac{1}{2}\|\mathbf{m}\cdot(x_i-\mathbf{D}\boldsymbol{\beta}_i)\|+\lambda\|\boldsymbol{\beta}_i\|_1$, where $\mathbf{m}$ is a mask vector to indicate whether a pixel is missing or not. The operator $\cdot$ is element-wise multiplication. The two methods produce almost identical results, as shown in Fig. 9. When using dictionaries with scale 85% of our estimated ones [1], the results are worse.

C. Human Action Recognition

Sparse dictionary learning was used in human action recognition [22]. We adopt the spatio-temporal interest point detector proposed by Dollar et al. [23]. The dense features are extracted following the procedures in [22] and [23]. Then samples of training and testing data are the extracted motion dense features in the video interest points. For those color dataset, we convert the RGB frame to gray scale one using [24].

1) Scale Adaption Evaluation: Our goal is to learn a human action dictionary from videos containing several actions. We use the KTH dataset [21], containing six types of human actions (walking, jogging, running, boxing, hand waving, and hand clapping) in the outdoor and indoor environment.

TABLE X
AVERAGE RECONSTRUCTION ERRORS FOR THE VIDEOS. THE CONFIGURATION IS THE SAME AS THAT IN TABLE V

Sparsity	Scale	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
10%	N	4.4024	4.3508	4.4226	4.3875	4.7437	4.6660
	$2N$	4.3828	4.2855	4.3596	4.3037	4.5059	4.5944
	$0.85N$	5.5022	5.2941	5.3583	4.9569	5.6816	5.5166
20%	N	3.9207	3.7306	3.7425	3.6357	4.0633	3.7027
	$2N$	3.8544	3.6324	3.5243	3.5579	3.9122	3.4536
	$0.85N$	4.5134	4.5975	4.4294	4.7406	4.7985	4.3366
40%	N	3.1611	3.3353	3.1249	3.3639	3.8236	3.5409
	$2N$	3.1581	3.2992	3.0244	3.2593	3.6750	3.3797
	$0.85N$	3.8711	4.1526	3.9902	4.2405	4.5636	3.9535

TABLE XI
ESTIMATED DICTIONARY SCALES ON THE SIX MAV SETS

MAV Set	1	2	3	4	5	6
Estimated $\lvert\mathbf{D}\rvert$	245	402	519	644	729	796

TABLE XII
ESTIMATED DICTIONARY SCALES BY BDL AND OUR METHOD ON THE SIX MAV SETS

MAV set	set 1	set 2	set 3	set 4	set 5	set 6
SADL	245	402	519	644	729	796
BDL	1067	1124	1349	1452	1556	1560

Traditionally, finding suitable dictionary scales need tryouts in this task.

A few examples are shown in Fig. 11, selected from the 589 short sequences. We index the actions from $1-6$. We define 6 mix-action video (MAV) sets similar to that for MTC: the i^{th} MAV set includes *all* data for actions $1 \sim i$. The estimated dictionary scales using our method on the 6 MAV sets are listed in Table XI. The tendency of increasing scales complies with our understanding of information richness in the input data. As the training set grows, scale increasing speed slows down. It is because the added patches share some common information with previous ones.

In quantitative evaluation, given learnt dictionary $\mathbf{D}$, we compare the average sparse reconstruction error $\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}-\mathbf{D}\boldsymbol{\beta}\|_2^2$ for 1000 randomly selected motion features, where $\boldsymbol{\beta}_i$ is the i^{th} training sample $\mathbf{x}_i$'s sparse coefficients over $\mathbf{D}$. The sparse reconstruction errors are listed in Table X. They also manifest that our results are complete and the estimated scales are suitable.

We again compare BDL with our SADL framework. The initial atom number for BDL is chosen as 2000 for MAV dictionaries. We tabulate the result in Table XII. Our method obtains more compact dictionaries.

2) Human Action Recognition: We implement the human action recognition framework of [22], [25]. Following the standard procedures, we compute an action descriptor $\mathbf{z}_i$ by the max pooling for the i^{th} video. $y_i \in \mathcal{Y} = \{1, \ldots, L\}$ is the label for the i^{th} data. Then given the training data $\{\mathbf{z}_i, \mathbf{y}_i\}$, a linear SVM is used to classify different human actions. We adopt the Leave-One-Out scheme. Our experiments are conducted on both the KTH dataset [21] and the Weizmann set [20]. The recognition accuracy using our learnt dictionaries, *safe dictionaries*, and *85%-dictionaries* is given in Table XIII. It is noticeable that the recognition accuracy using our dictionary and the safe one is very close. When the 85%-dictionaries are used, the rates drop.

TABLE XIII
RECOGNITION RATES

	safe dictionary	85%-dictionary	ours
KTH	93.08%	87.83%	93.17%
Weizmann	92.10%	86.71%	92.05%

The detailed recognition accuracies on our learned dictionaries, "safe dictionary" and "85%-dictionary" are illustrate in Fig. 12, represented by confusion matrices. Confusion matrix is widely used in human action recognition to evaluate results considering different actions. The element in the i^{th} row and j^{th} column means the percentage of action i being classified into action j. A good result is expected to have diagonal elements close to 1.

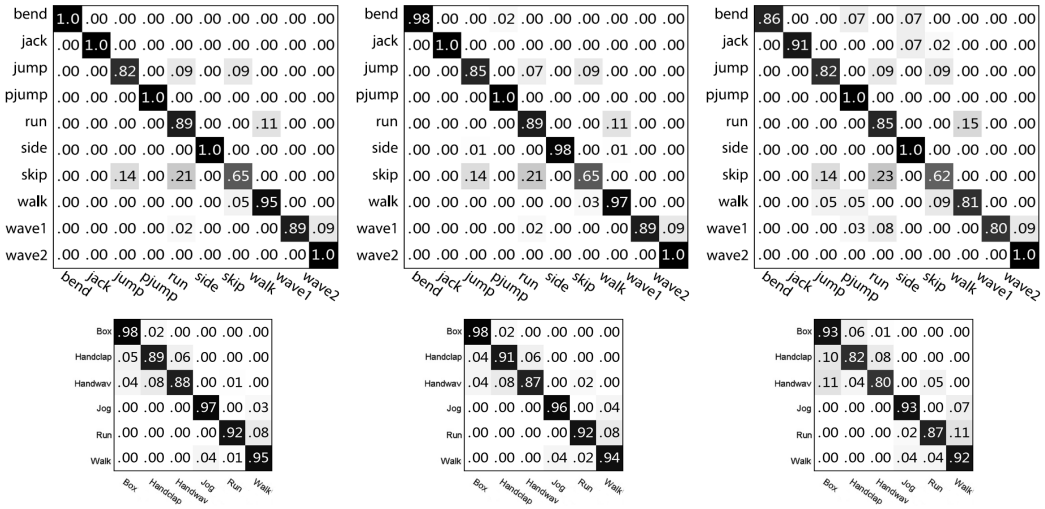
D. Unusual Event Detection

We also demonstrate that unusual event detection can benefit from our SADL method. This task needs to learn normal event patterns. Then any incoming frame that is greatly deviated from these normal patterns is labeled as unusual. Learning a dictionary for each local region is a common choice. In our experiment, we resize each frame to 120×160 pixels with 12×16 regular patches. So each patch is with 10×10 pixels.

Obviously, scales of dictionaries in different regions cannot be identical, since normal event patterns vary from region to region. We extract motion features following that of [26] and learn a dictionary for each subregion. Unusual subregions are those with the reconstruction error larger than a threshold (0.2 in our experiments). When the number of unusual subregions in 3 consecutive frame exceeds 30, an unusual event is detected. We test our method on two datasets, i.e., UCSD Ped1 dataset [27] and Subway dataset [28].

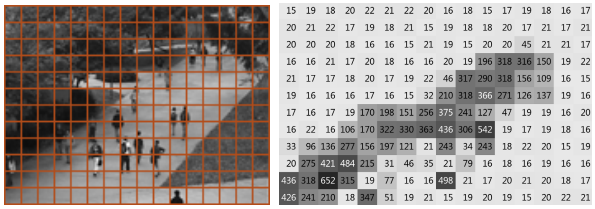
To demonstrate the scale adaption ability of our SADL method, we report our learnt dictionary scales in different subregions on the UCSD dataset in Fig. 13. In regions containing tree structures, the motion pattern is mostly regular. Therefore, a small dictionary is enough. On the contrary, the road regions involve complex crowd motion, which requires large dictionaries.

We compare our SADL with traditional dictionary learning [9] that sets the same scale for all dictionaries for different subregions. For fairness, we test setting a variety of scales including 50, 100, 200, 400, and 800 for the dictionaries. We report the results on the Subway dataset in Table XIV. With automatic dictionary scale estimation, our method runs faster and yields the more accurate detection result. We also compare results on the UCSD Ped1 Dataset.



(a) (b) (c)

Fig. 12. Recognition accuracy represented using confusion matrices. The first row is for the result in the KTH dataset; the second row shows the result in the Weizmann dataset. In each confusion matrix, x-coordinate indexes ground truth action and y-coordinates are for different actions.(a) Our method. (b) "Safe dictionary." (c) "85%-dictionary."



(a) (b)

Fig. 13. (a) Frame in the UCSD Ped1 Dataset. (b) Reports learnt dictionary scales in different subregions.

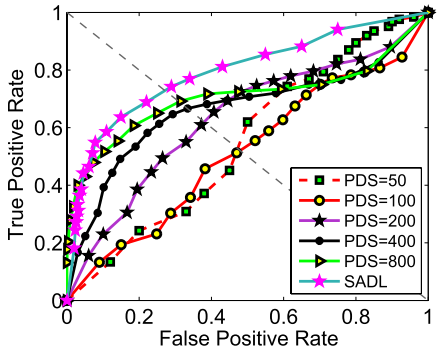


Fig. 14. ROC curve on the UCSD Ped1 Dataset [27].

TABLE XIV

RESULT COMPARISON ON THE SUBWAY ENTRANCE VIDEO. "GT" STANDS FOR GROUND TRUTH. "PDS" MEANS ALL SUBREGIONS HAVE THE SAME DICTIONARY SCALE. EVENTS INCLUDE WD (WRONG DIRECTION), NP (NO PAYMENT), LT (LOITERING), II (IRREGULAR INTERACTIONS), ALL (SUM OF ALL UNUSUAL CASES), AND FA (FALSE ALARM)

	WD	NP	LT	II	misc	All	FA
GT	26	13	14	4	9	66	0
PDS = 50	20	7	12	4	7	50	27
PDS = 100	22	7	12	4	7	52	19
PDS = 200	21	8	11	4	7	51	11
PDS = 400	20	6	11	4	8	49	6
PDS = 800	20	7	10	4	7	48	5
Ours	23	9	12	4	8	**56**	**5**

We tune the threshold (number of unusual subregions) to plot the ROC curve, given in Fig. 14.

These experiments show that scale adaptation for dictionary learning is important. If the assigned scale is lower than necessary, normal patterns may not be well represented, resulting in more false alarms. On other hand, an overly large dictionary may smoothly represent abnormal patterns, increasing ambiguity. Note that hand-tuning these scales for all regions is impossible.

VI. CONCLUSION

We have presented a new model to automatically estimate the dictionary size during learning. It involves Atom Indicator Vectors (AIVs) to indicate if one basis is important or not by evaluating the responses. The final function is solved by approximating the novel dimension constraining term by a Multivariate Moreau Proximal Indicator (MMPI) penalty. We evaluate the effectiveness of our system using texture and human action examples. They indicate that our estimated dictionary scale is suitable. Our framework is general. It could possibly benefit many image processing and computer vision problems and helps save time and effort in finding correct scales.

APPENDIX

Proof of Theorem 1: Assuming $\{\mathbf{D}^*, \mathcal{A}^*\}$ is the optimal solution of Eq. (3), any two atoms $\mathbf{d}_v^*$ and $\mathbf{d}_u^*$ with $\mathbf{I}(\widehat{\boldsymbol{\alpha}}_v^*) = 1$ and $\mathbf{I}(\widehat{\boldsymbol{\alpha}}_u^*) = 1$ must satisfy

$$\|\mathbf{d}_v^* - \mathbf{d}_u^*\|_2^2 \geq \frac{n\mu\lambda^2}{\kappa\phi^2}, \tag{20}$$

where $\phi = \sum_{i=1}^{n}\{\frac{1}{2}\|\mathbf{x}_i\|_2^2\}$ and $\kappa = 1 + \frac{\lambda}{\sqrt{n\phi}}$.

Proof: The objective function for our model in Eq. (18) consists of the following three parts, i.e.,

$$\begin{aligned} E_1(\mathbf{D},\mathcal{A}) &= \frac{1}{n}\sum_{i=1}^{n}\{\frac{1}{2}\|\mathbf{D}\boldsymbol{\alpha}_i-\mathbf{x}_i\|_2^2\},\\ E_2(\mathbf{D},\mathcal{A}) &= \lambda\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{\alpha}_i\|_1,\\ E_3(\mathbf{D},\mathcal{A}) &= \mu\sum_{j=1}^{k}\mathbf{I}(\widehat{\boldsymbol{\alpha}}_j). \end{aligned} \tag{21}$$

Here, $E_1(\mathbf{D},\mathcal{A})$ is the data fitting term, $E_2(\mathcal{A},\mathbf{D})$ is the sparse-inducing term, and $E_3(\mathcal{A},\mathbf{D})$ is the dictionary scale penalty term.

Suppose $\{\mathbf{D}^*,\mathcal{A}^*\}$ is an optimal solution. We can construct another solution $\{\mathbf{D}^*,\mathcal{A}^+\}$ as

$$\alpha_{i,j}^{+}=\begin{cases}\alpha_{i,\upsilon}^{*}+\alpha_{i,u}^{*} & \text{if } j=u\\ 0 & \text{if } j=\upsilon\\ \alpha_{i,j}^{*} & \text{otherwise}\end{cases}\quad\forall i=1,\ldots,n. \tag{22}$$

We discuss the relationship between $\{\mathbf{D}^*,\mathcal{A}^*\}$ and $\{\mathbf{D}^*,\mathcal{A}^+\}$ with regard to the above three terms respectively. For the data fitting term, we have

$$E_1(\mathbf{D}^*,\mathcal{A}^*)=\frac{1}{2n}\sum_{i=1}^{n}\{\|\mathbf{R}_i+\alpha_{i,\upsilon}^{*}\mathbf{d}_\upsilon^{*}+\alpha_{i,u}^{*}\mathbf{d}_u^{*}\|_2^2\}, \tag{23}$$

where $\mathbf{R}_i=\sum_{j\in\Omega}\alpha_{i,j}^{*}\mathbf{d}_j^{*}-\mathbf{x}_i$, and Ω is the index set excluding u and υ. Moreover, Eq. (23) is equivalent to

$$\begin{aligned} &E_1(\mathbf{D}^*,\mathcal{A}^*)\\ &=\frac{1}{2n}\sum_{i=1}^{n}\{\|\mathbf{R}_i+(\alpha_{i,\upsilon}^{*}+\alpha_{i,u}^{*})\mathbf{d}_\upsilon^{*}+\alpha_{i,u}^{*}(\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*})\|_2^2\}. \end{aligned} \tag{24}$$

Given vectors $\mathbf{a}$ and $\mathbf{b}$, the following inequality holds.

$$\|\mathbf{a}+\mathbf{b}\|_2^2\geq\|\mathbf{a}\|_2^2-\|\mathbf{b}\|_2^2-2\|\mathbf{a}+\mathbf{b}\|_2\|\mathbf{b}\|_2. \tag{25}$$

We can derive an inequality from the above three equations, written as

$$\begin{aligned} E_1(\mathbf{D}^*,\mathcal{A}^*)\geq\ & E_1(\mathbf{D}^*,\mathcal{A}^+)-\frac{1}{2n}\|\widehat{\boldsymbol{\alpha}}_u^{*}\|_2^2\|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2^2\\ &-\frac{1}{n}\sum_{i=1}^{n}|a_{i,u}^{*}|\cdot\sqrt{E_1(\mathbf{D}^*,\mathcal{A}^*)}\cdot\|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2. \end{aligned} \tag{26}$$

Given $\|\cdot\|_1$ an element-wise operator, the sparse term can be written as

$$\begin{aligned} E_2(\mathbf{D}^*,\mathcal{A}^*)=\ & \lambda\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{\alpha}_i^{*}\|_1=\lambda\frac{1}{n}\sum_{j=1}^{k}\|\widehat{\boldsymbol{\alpha}}_j^{*}\|_1\\ =\ & \lambda\frac{1}{n}\{\sum_{j\in\Omega}\|\widehat{\boldsymbol{\alpha}}_j\|_1+\|\widehat{\boldsymbol{\alpha}}_u+\widehat{\boldsymbol{\alpha}}_\upsilon\|_1\\ &+\|\widehat{\boldsymbol{\alpha}}_u\|_1+\|\widehat{\boldsymbol{\alpha}}_\upsilon\|_1-\|\widehat{\boldsymbol{\alpha}}_u+\widehat{\boldsymbol{\alpha}}_\upsilon\|_1\}\\ =\ & E_2(\mathbf{D}^*,\mathcal{A}^+)+\lambda\frac{1}{n}\{\|\widehat{\boldsymbol{\alpha}}_u\|_1+\|\widehat{\boldsymbol{\alpha}}_\upsilon\|_1\\ &-\|\widehat{\boldsymbol{\alpha}}_u+\widehat{\boldsymbol{\alpha}}_\upsilon\|_1\}\geq E_2(\mathbf{D}^*,\mathcal{A}^+) \end{aligned} \tag{27}$$

since the norm operator satisfies the triangle inequality

$$\|\widehat{\boldsymbol{\alpha}}_u\|_1+\|\widehat{\boldsymbol{\alpha}}_\upsilon\|_1\geq\|\widehat{\boldsymbol{\alpha}}_u+\widehat{\boldsymbol{\alpha}}_\upsilon\|_1. \tag{28}$$

For the dictionary size penalty term, we have

$$\sum_{j=1}^{k}\mathbf{I}(\widehat{\boldsymbol{\alpha}}_j)=\sum_{j\in\Omega}\mathbf{I}(\widehat{\boldsymbol{\alpha}}_j)+\mathbf{I}(\widehat{\boldsymbol{\alpha}}_\upsilon+\widehat{\boldsymbol{\alpha}}_u)+\mathbf{I}(\widehat{\boldsymbol{\alpha}}_\upsilon)+\mathbf{I}(\widehat{\boldsymbol{\alpha}}_u)-\mathbf{I}(\widehat{\boldsymbol{\alpha}}_\upsilon+\widehat{\boldsymbol{\alpha}}_u).$$

We thus can write

$$\begin{aligned} &E_3(\mathbf{D}^*,\mathcal{A}^*)\\ &\quad= E_3(\mathbf{D}^*,\mathcal{A}^+)+\mu(\mathbf{I}(\widehat{\boldsymbol{\alpha}}_\upsilon)+\mathbf{I}(\widehat{\boldsymbol{\alpha}}_u)-\mathbf{I}(\widehat{\boldsymbol{\alpha}}_\upsilon+\widehat{\boldsymbol{\alpha}}_u))\\ &\quad\geq E_3(\mathbf{D}^*,\mathcal{A}^+)+\mu. \end{aligned} \tag{29}$$

since $\mathbf{d}_u^{*}$ and $\mathbf{d}_\upsilon^{*}$ are selected as output dictionary atoms, it is natural that $\mathbf{I}(\widehat{\boldsymbol{\alpha}}_u)=\mathbf{I}(\widehat{\boldsymbol{\alpha}}_\upsilon)=1$, making

$$\mathbf{I}(\widehat{\boldsymbol{\alpha}}_\upsilon)+\mathbf{I}(\widehat{\boldsymbol{\alpha}}_u)-\mathbf{I}(\widehat{\boldsymbol{\alpha}}_\upsilon+\widehat{\boldsymbol{\alpha}}_u)\geq 1. \tag{30}$$

Combining Eqs. 26, 27 and 29, we obtain

$$\begin{aligned} \mathbf{E}(\mathbf{D}^*,\mathcal{A}^*)\geq\ & \mathbf{E}(\mathbf{D}^*,\mathcal{A}^+)+\mu-\frac{1}{2n}\|\widehat{\boldsymbol{\alpha}}_u^{*}\|_2^2\|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2^2\\ &-\frac{1}{n}\sum_{i=1}^{n}|a_{i,u}^{*}|\cdot\sqrt{E_1(\mathbf{D}^*,\mathcal{A}^*)}\cdot\|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2. \end{aligned} \tag{31}$$

As $\{\mathcal{A}^*,\mathbf{D}^*\}$ is the optimum solution, it holds that $\mathbf{E}(\mathbf{D}^*,\mathcal{A}^*)-\mathbf{E}(\mathbf{D}^*,\mathcal{A}^+)\leq 0$. Hence, we have

$$\|\widehat{\boldsymbol{\alpha}}_u^{*}\|_2^2\|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2^2+2\|\widehat{\boldsymbol{\alpha}}_u^{*}\|_1\sqrt{E_1(\mathbf{D}^*,\mathcal{A}^*)}\|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2\geq 2n\mu. \tag{32}$$

Eq. (32) involves both $\|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2^2$ and $\|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2$. We simplify it by computing the upper bound of $\|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2$ as

$$\begin{aligned} &\|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2\leq\|\mathbf{d}_u^{*}\|_2+\|\mathbf{d}_\upsilon^{*}\|_2=2\\ \Longrightarrow\ &\|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2^2\leq 2\|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2 \end{aligned} \tag{33}$$

Therefore, Eq. (32) can be reformulated as

$$(\|\widehat{\boldsymbol{\alpha}}_u^{*}\|_2^2+\sqrt{E_1(\mathbf{D}^*,\mathcal{A}^*)}\|\widehat{\boldsymbol{\alpha}}_u^{*}\|_1)\ \|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2\geq n\mu. \tag{34}$$

What we need to do now is to estimate upper bounds of $\|\widehat{\boldsymbol{\alpha}}_u^{*}\|_2^2$, $\|\widehat{\boldsymbol{\alpha}}_u^{*}\|_1$ and $E_1(\mathbf{D}^*,\mathcal{A}^*)$. Further, there are two inequalities expressed as

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{2}\|\mathbf{x}_i\|_2^2\right\}&=\mathbf{E}(\mathbf{D}^*,\mathbf{0})\geq\mathbf{E}(\mathbf{D}^*,\mathcal{A}^*)\\ &\geq E_2(\mathbf{D}^*,\mathcal{A}^*)\geq\frac{\lambda}{n}\|\boldsymbol{\alpha}_u^{*}\|_1\geq\frac{\lambda}{n}\|\widehat{\boldsymbol{\alpha}}_u^{*}\|_2, \end{aligned} \tag{35}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{2}\|\mathbf{x}_i\|_2^2\right\}=\mathbf{E}(\mathbf{D}^*,\mathbf{0})\geq\mathbf{E}(\mathbf{D}^*,\mathcal{A}^*)\geq E_1(\mathbf{D}^*,\mathcal{A}^*), \tag{36}$$

where $\mathbf{0}$ is the matrix whose elements are all zeros. Combining Eqs. 32, 35 and 36, we get

$$(\frac{\phi^2}{\lambda^2}+\frac{\phi}{\lambda}\sqrt{\frac{\phi}{n}})\|\mathbf{d}_u^{*}-\mathbf{d}_\upsilon^{*}\|_2\geq n\mu, \tag{37}$$

where $\phi = \sum_{i=1}^{n} \{\frac{1}{2}\|\mathbf{x}_i\|_2^2\}$. It further leads to

$$\|\mathbf{d}_u^* - \mathbf{d}_v^*\|_2 \geq \frac{\kappa n \lambda^2}{\phi^2}\mu. \tag{38}$$

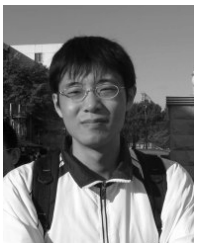
where $\phi = \sum_{i=1}^{n} \{\frac{1}{2}\|\mathbf{x}_i\|_2^2\}$ and $\kappa = 1/(1 + \frac{\lambda}{\sqrt{n\phi}})$. ■

ACKNOWLEDGMENT

This work is supported by a grant from the Research Grants Council of the Hong Kong SAR (project No. 413110) and by NSF of China (key project No. 61133009).

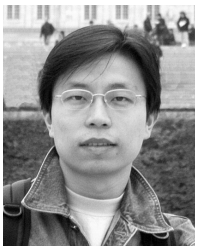
REFERENCES

[1] E. Michael and A. Micha, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Image Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[2] G. Peyré, "Sparse modeling of textures," *J. Math. Imag. Vis.*, vol. 34, no. 1, pp. 17–31, 2009.
[3] T. Ivana and F. Pascal, "Dictionary learning for stereo image representation," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 921–934, Apr. 2011.
[4] E. Michael and A. Michal, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.
[5] J. Shi, X. Ren, G. Dai, J. Wang, and Z. Zhang, "A non-convex relaxation approach to sparse dictionary learning," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 1809–1816.
[6] Y. Jianchao, W. John, H. Thomas, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.
[7] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. ICCV*, 2013.
[8] C. Lu, J. Shi, and J. Jia, "Online robust dictionary learning," in *Proc. IEEE Conf. CVPR*, 2013.
[9] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 19–60, 2010.
[10] M. Zhou, H. Chen, P. John, L. Ren, S. Guillermo, and C. Lawrence, "Non-parametric Bayesian dictionary learning for sparse image representations," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2009.
[11] I. Ramirez and G. Sapiro, "An MDL framework for sparse coding and dictionary learning," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2913–2927, Jun. 2012.
[12] B. Christopher, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.
[13] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Image Process.*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.
[14] 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.
[15] A. Pierre, M. Olivier, and R. Tyrrell, *Nonsmooth Mechanics and Analysis: Theoretical and Numerical Advances*, vol. 12. New York, NY, USA: Springer-Verlag, 2006.
[16] A. Michal and E. Michael, "Sparse and redundant modeling of image content using an image-signature-dictionary," *SIAM J. Imag. Sci.*, vol. 1, no. 3, pp. 228–247, 2008.
[17] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
[18] C. Antonio, P. Patrick, and T. Kentaro, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.
[19] P. Nikos and D. Rachid, "Geodesic active regions and level set methods for supervised texture segmentation," *Int. J. Comput. Vis.*, vol. 46, no. 3, pp. 223–247, 2002.
[20] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. ICCV*, vol. 2. 2005, pp. 1395–1402.
[21] S. Christian, L. Ivan, and C. Barbara, "Recognizing human actions: A local SVM approach," in *Proc. 17th ICPR*, Aug. 2004, pp. 32–36.
[22] C. Liu, Y. Yang, and Y. Chen, "Constructing visual vocabularies using sparse coding for action recognition," in *Proc. ICIECS*, Dec. 2009, pp. 1–4.
[23] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Vis. Surveill. Perform. Evaluat. Track.*, Oct. 2005, pp. 65–72.
[24] C. Lu, L. Xu, and J. Jia, "Contrast preserving decolorization," in *Proc. IEEE ICCP*, Apr. 2012, pp. 1–7.
[25] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1794–1801.
[26] N. J. Carlos, H. Wang, and F.-F. Li, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.*, vol. 79, no. 3, pp. 299–318, 2008.
[27] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 935–942.
[28] A. Amit, R. Ehud, S. Ilan, and R. Daviv, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.



Cewu Lu received the BS and MS degrees from Chongqing University of Posts and Telecommunications and Graduate University of Chinese Academy of Sciences in 2006 and 2009 respectively, and the PhD degree in 2013 in computer science and engineering from the Chinese University of Hong Kong. He is currently a research fellow at the Hong Kong University of Science and Technology. He received of the best paper award of NPAR 2012 and served as reviewers for several major computer vision and graphics conferences and journals. His research interests include activity recognition, dictionary learning, image/video processing. He is a member of the IEEE.

Jianping Shi received the BS degree in computer science and engineering from Zhejiang University, China in 2011. She is now pursuing her PhD degree in the Chinese University of Hong Kong. She received the Hong Kong PhD Fellowship and Microsoft Research Asia Fellowship Award in 2011 and 2013 respectively. Her research interests include in computer vision and machine learning. She is a member of the IEEE.



Jiaya Jia received the PhD degree in Computer Science from Hong Kong University of Science and Technology in 2004 and is currently an associate professor in Department of Computer Science and Engineering at the Chinese University of Hong Kong (CUHK). He was a visiting scholar at Microsoft Research Asia from March 2004 to August 2005 and conducted collaborative research at Adobe Systems in 2007. He heads the research group in CUHK, focusing specifically on computational photography, 3D reconstruction, practical optimization, and motion estimation. He currently serves as an associate editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) and served as an area chair for ICCV 2011 and ICCV 2013. He was on the program committees of several major conferences, including ICCV, ECCV, ICCP, and CVPR, and co-chaired the Workshop on Interactive Computer Vision, in conjunction with ICCV 2007. He received the Young Researcher Award 2008 and Research Excellence Award 2009 from CUHK. He is a senior member of the IEEE.