Multimed Tools Appl (2016) 75:1409–1425
DOI 10.1007/s11042-014-2179-8

Visual concept detection of web images based on group sparse ensemble learning

Yongqing Sun · Kyoko Sudo · Yukinobu Taniguchi

Received: 28 December 2013 / Revised: 1 July 2014 / Accepted: 3 July 2014 /
Published online: 25 July 2014

Abstract Due to the huge intra-class variations for visual concept detection, it is necessary for concept learning to collect large scale training data to cover a wide variety of samples as much as possible. But it presents great challenges on both how to collect and how to train the large scale data. In this paper, we propose a novel web image sampling approach and a novel group sparse ensemble learning approach to tackle these two challenging problems respectively. For data collection, in order to alleviate manual labeling efforts, we propose a web image sampling approach based on dictionary coherence to select coherent positive samples from web images. We propose to measure the coherence in terms of how dictionary atoms are shared because shared atoms represent common features with regard to a given concept and are robust to occlusion and corruption. For efficient training of large scale data, in order to exploit the hidden group structures of data, we propose a novel group sparse ensemble learning approach based on Automatic Group Sparse Coding (AutoGSC). After AutoGSC, we present an algorithm to use the reconstruction errors of data instances to calculate the ensemble gating function for ensemble construction and fusion. Experiments show that our proposed methods can achieve promising results and outperforms existing approaches.

Keywords Ensemble learning · Visual concept detection · Semantic indexing · Web image mining · Sparse representation · Dictionary learning

The preliminary version of this paper was partly published in the Pacific-Rim Conference on Multimedia (PCM 2013), and partly in the 19th International Conference on Multimedia Modeling (MMM 2013).

Y. Sun (✉) · Y. Sudo · Y. Taniguchi
NTT Media Intelligence Laboratories
1-1 Hikarinooka Yokosuka-Shi Kanagawa 239-0847, Japan
e-mail: yongqing.sun@lab.ntt.co.jp

Y. Sudo
e-mail: sudo.kyoko@lab.ntt.co.jp

Y. Taniguchi
e-mail: taniguchi.yukinobu@lab.ntt.co.jp

1 Introduction

With the advent of the big data era, the explosive growth of visual contents on the Internet presents a challenge in how to manage the ever-growing size of the multimedia collections, particularly in how to extract sufficiently accurate semantic metadata (concepts) to make them searchable [9, 11, 25]. Visual concept detection is essentially a classification task in which classifiers are learned with various features extracted from training samples to predict the presence of a certain concept in a video shot or keyframe (image) [25, 29]. Ranging from objects such as *"boat"* and *"car"* to scenes such as *"sky"* and *"sea"*, semantic concepts can serve as good intermediate semantic metadata for video content indexing and understanding [25]. Establishing a large set of robust concept detectors will yield significant improvements in many challenging applications, such as image/video search and summarization [29].

Due to the existence of the well-known semantic gap [18] between the low level visual features and the users's semantic interpretation of diversified visual data, concept detection is a challenging yet essential task that has attracted the attention of many researchers [25]. The visual content for a given concept often possess huge variations resulting from diversified visual appearances, camera shooting and video editing styles, etc. Such huge intra-class variations hinders the performance of most machine learning approaches [25].

To solve the problem of huge intra-class variations, it may be a promising solution to collect large scale training data to cover a wide variety of samples as much as possible. Previous studies on visual concept detection [9] and pedestrian classification [8, 14] indicate that the data matters most; this was highlighted recently by machine learning researchers who stated that "the quickest path to success is often to just get more data, and more data beats a cleverer algorithm" [7]. In order to learn effective concept detectors, a critical step is to acquire a sufficiently large amount of training samples, especially positive training samples [9]. However, how to collect and label large scale training data is very challenging since the data collection and manual labeling are laborious and time consuming. Fortunately, with the explosive growth of visual contents on the Internet, large amounts of training samples have become available through Web searching [5, 29]. Consequently, how to utilize these abundant web images to improve concept detection has been the subject of intensive research by a large multimedia research community, since it has offered promising ways to automatically annotate the contents at relatively low cost [5, 29].

Furthermore, with the increasing of training dataset size, the training may be very time consuming since the time complexity of most machine learning methods such as Support Vector Machine (SVM) is between $O\left(n^2\right)$ and $O\left(n^3\right)$ (n is the number of training samples) [4, 25]. This seems infeasible if the number of training samples is very large, such as over one hundred thousand, and the feature dimension is very high. Therefore, for large scale dataset, how to train it effectively and efficiently is also a big challenge.

In this paper, we propose to an novel web image sampling approach and a novel group sparse ensemble learning approach to tackle these two challenging problems respectively. For data collection, in order to alleviate manual labeling effort, we propose a web image sampling approach based on dictionary coherence to select coherent positive samples from web images based on the degree of image coherence with a given concept. We propose to measure the coherence in terms of how dictionary atoms are shared since shared atoms represent common features with regard to a given concept and are robust to occlusion and corruption. Thus, two kinds of dictionaries are learned through online dictionary learning

methods: one is the concept dictionary learned from key-point features of all the positive training samples while the other is the image dictionary learned from those of web images. Intuitively, the coherence degree is then calculated by the Frobenius norm of the product matrix of the two dictionaries. For efficient training of large scale data, in order to exploit the hidden group structures of data, we propose a novel group sparse ensemble learning approach based on Automatic Group Sparse Coding (AutoGSC). We first adopt AutoGSC to learn both a common dictionary over different data groups and an individual group-specific dictionary for each data group which can help us to capture the discrimination information contained in different data groups. Next, we represent each data instance by using a sparse linear combination of both dictionaries. Finally, we propose an algorithm to use the reconstruction errors of data instances to calculate the ensemble gating function for ensemble construction and fusion.

The main contribution of this paper is that we propose a novel web image sampling approach for training data collection and a novel group sparse ensemble learning approach for efficient visual concept detection.The rest of the paper is organized as follows. We first review the related work on web image sampling and visual concept Learning respectively in Section 2. Then we describe our proposed web image sampling approach based on dictionary coherence and our group sparse ensemble learning method respectively in Section 3. We will describe our experiments and give our experimental results in Section 4. Finally, we will conclude our work in Section 5.

2 Related work

2.1 Web image sampling

As aforementioned, how to utilize web images to improve concept detection has been the subject of intensive research by a large multimedia research community due to its relatively low cost for manual annotation [5, 29]. [29] empirically studied the effect of exploiting tagged images on concept learning by analyzing tag lists. [5] proposed an automatic concept-to-query mapping method for acquiring training data from online platforms.

However, the online web images are very noisy, often cover a wide range of unpredictable contents, and have quite different data distributions with any close dataset such as TRECVid dataset [15, 22]. As shown in Fig. 1, for example, the content of web images searched from Google Image with the keyword “Airplane-flying” varies greatly. Obviously, the images in the top row of the figure are incoherent from the concept “Airplane-flying” in the TRECVid dataset. Thus these images can not facilitate the training of the concept and may even harm it. Only the images in the second row are consistent with the dataset and hence helpful. Therefore, how to select coherent positive training samples from diffused web images is a challenging problem for training of effective concept detectors [5, 20, 21] due to the existence of cross-domain incoherence resulting from the mismatch of data distributions.

Existing work on video concept learning using web images has mainly focused on how to leverage compact features, such as region-based features [21] or image salience [20], to alleviate the visual differences. Since an image is greatly reduced to a very compact feature vector, the effect of these approaches is not evident. In this paper, we propose a novel sampling approach on how to exploit bundles of local key-point features to measure how coherent a web image is with a given concept, from the aspects of sparse coding and dictionary learning.

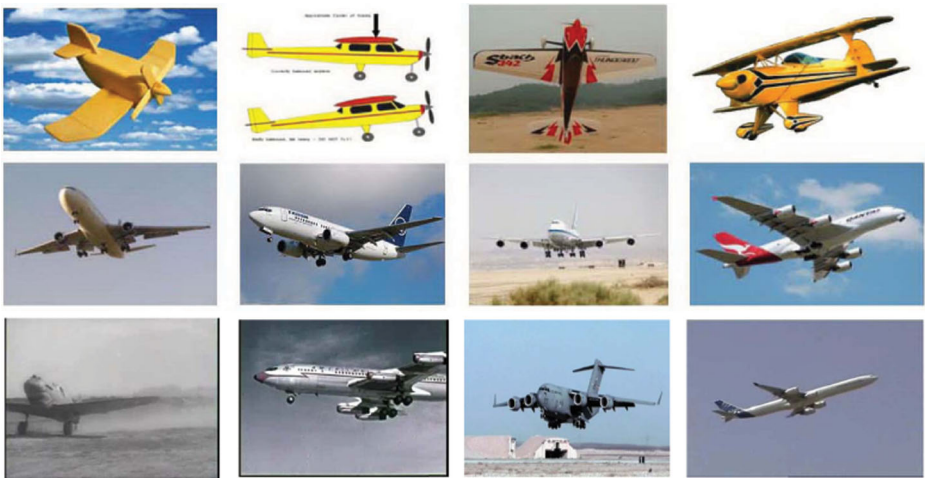

Fig. 1 Web Image Examples of "Airplane-flying" (the first two rows) compared with the positive examples of "Airplane-flying" in the TRECVID 2012 semantic indexing task (the last row)

2.2 Visual concept learning

Due to the low efficiency and unscalability of the classical methods based on global classification such as SVM [6], Gaussian Mixture Model [1], Hidden Markov Model [16], statistical active learning[27], and various ensemble learning methods such as LDA-SVM [22–24], multi-kernel ensemble learning [19] and sparse ensemble learning [25], were developed for visual concept detection; they exploit the "divide and conquer" strategy to train large amounts of samples both effectively and efficiently.

In particular, [25] proposed an efficient sparse ensemble learning method by exploiting a sparse non-negative matrix factorization process for ensemble construction and fusion. It was shown to achieve state-of-art performance on the TRECVid 2008 dataset. However, this approach adopts traditional sparse coding and so treats each data instance as an individual and no data group information is considered. It considers each visual feature such as Bag of Visual Word (BoVW) of an image as a separate coding problem and does not take into account the fact that the sparse coding of each feature does not guarantee the sparse coding of all images in the dataset.

Each dataset usually consists of many categories, and is assured of having hidden group structures [28]. Once a dictionary atom has been selected to represent an image, it may as well as be used to represent other images of a given category without much additional regularization cost [3]. Therefore, it makes more sense to learn a group level sparse representation [3]. To exploit the group structures hidden in the data set, [3] proposed Group Sparse Coding (GSC), which learns a sparse representation on the group level as well as a shared dictionary. However, GSC assumes that the data group identities are pre-given, even though they are often hidden in many real world applications, and it can only learn a common dictionary [26]. To discover the hidden data groups, [26] proposed Automatic Group Sparse Coding (AutoGSC) by learning both a common dictionary over different data groups and an individual group-specific dictionary for each data group which can help us to capture the discrimination information contained in different data groups.

Inspired by the sparse ensemble learning work [25] and the advantages of AutoGSC [26] in discovering hidden structures of data, in this paper, we propose a novel group sparse ensemble learning approach based on automatic group sparse coding to exploit the hidden group structures of data.

3 Proposed approaches

Due to the huge intra-class variations for visual concept detection, it is necessary for concept learning to collect large scale training data to cover a wide variety of samples as much as possible. But it presents both great challenges on both how to collect and how to train the large scale data. In this section, we will elaborate on the details of our proposed web image sampling approach and group sparse ensemble learning approach to tackle these two challenging problems respectively.

3.1 Web image sampling

3.1.1 Overview

Inspired by the observation that dictionary atoms representing common features in all categories tend to appear to be repeated almost exactly in dictionaries corresponding to different categories, [17] promotes incoherence between the dictionary atoms to improve the speed and accuracy of sparse coding.

Motivated by this work, since the shared dictionary atoms learned from data can represent common features with regard to a given concept (represented by the set of positive training samples) and are robust to occlusion and corruption [13], we propose to use dictionary coherence in terms of how an image and a given concept share dictionary atoms to measure the degree of image coherence with the concept. That is, the more atoms they share, the higher the dictionary coherence is, which means it is more probable that the web image is coherent with the concept.

In order to compute the dictionary coherence, we learn two kinds of dictionaries through the online dictionary learning method [13]: one is the concept dictionary learned from key-point features of all the positive training samples while the other is the image dictionary learned from those of web images. Intuitively, the coherence degree is then calculated by the Frobenius norm of the product matrix of the two dictionaries since it reflects the sum of the absolute values of inner products between dictionary atoms.

On the basis of the dictionary coherence, we propose a novel adaptive sampling approach to select coherent positive samples from diffused web images for further concept learning.

3.1.2 Algorithm

As shown in the framework of Fig. 2, for each concept, the algorithm of the proposed sampling principally consists of the following steps:

(1) Construction of concept set: Select all the positive training samples from a development dataset such as TRECVid development set to represent the concept.

(2) Feature extraction of concept set: Extract local key-point features, such as SIFT [12] or SURF [2], and collect each key-point feature $x_i \in \mathbf{R}^d$ of all the images in the

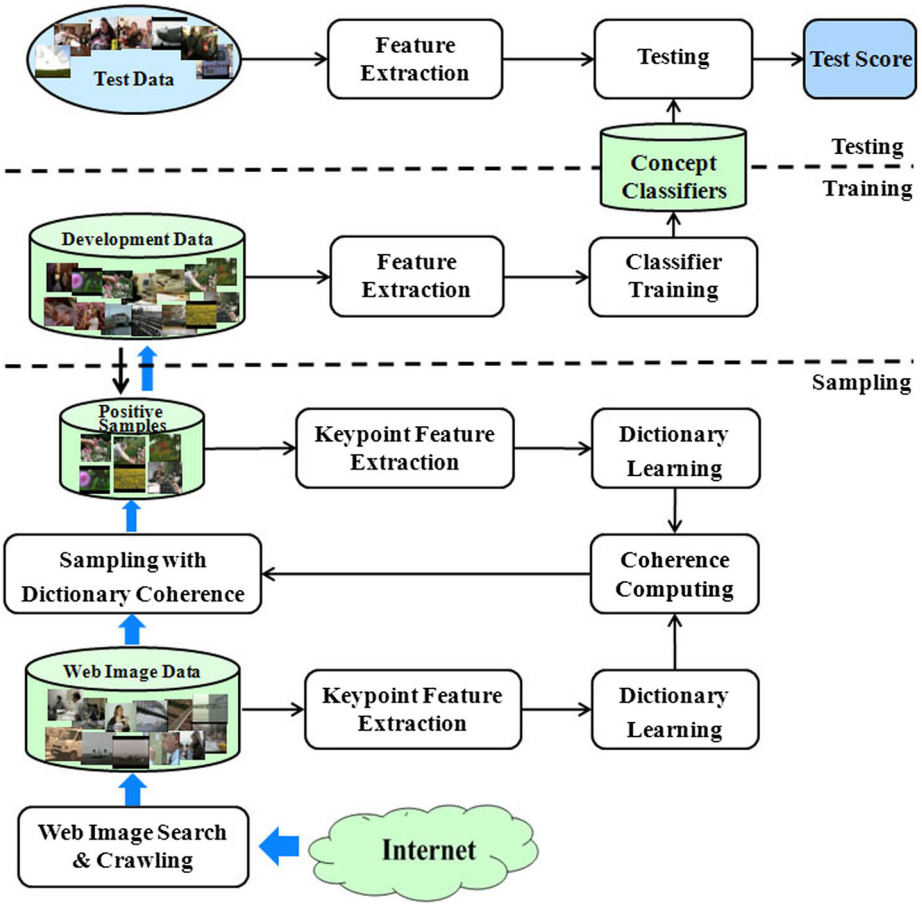

Fig. 2 Proposed Framework of Web Image Sampling

concept set to form the data matrix $\mathbf{X}_c = [x_1, \ldots, x_n] \in \mathbf{R}^{d \times n}$. Here, d is the feature dimensionality, and n is the total number of keypoints.

(3) Concept dictionary learning: Adopt the efficient online dictionary learning methods [13] to learn the concept dictionary $\mathbf{D}_c \in \mathbf{R}^{d \times k}$ from the concept data matrix $\mathbf{X}_c$, where k is the size of the dictionary, i.e., the number of atoms. For the SIFT feature, we set $k = 192$ about 1.5 to 2.0 times of the feature size $d = 128$ [25].

(4) Collection of web image set: After query construction or mapping [5] based on the concept name, search the web images and crawl the top-ranked ones.

(5) Feature extraction of web image: For each image in the web image set, extract the same local key-point features as the second step, and form the image data matrix $\mathbf{X}_i \in \mathbf{R}^{d \times m}$, where m is the number of keypoints in the image.

(6) Image dictionary learning: Adopt the same dictionary learning methods [13] to learn the image dictionary $\mathbf{D}_i \in \mathbf{R}^{d \times k}$ from the image data matrix $\mathbf{X}_i$.

(7) Dictionary coherence computing: Use (2) in subsection 3.1.4 to compute the dictionary coherence C_i between the image dictionary $\mathbf{D}_i$ and the concept dictionary $\mathbf{D}_c$.

(8) Adaptive sampling: Compare the dictionary coherence C_i of the current web image with the adaptive threshold in subsection 3.1.5 to determine whether to add the current web image to the training set.

As shown in Fig. 2, after adding the selected coherent positive web samples (a manual check is advised to ensure it is positive) to the training set, we can do further concept learning for training more effective concept detectors. We will detail the key procedures in the following subsections.

3.1.3 Dictionary learning

In our study, we use the efficient online learning methods [13] to learn the dictionary. Due to the advantage of non-negativity constraints in learning part-based representations [25], which is helpful for object-oriented concept learning, we impose the positivity constraints on both dictionary D and sparse code α_i in solving the optimization problem as below:

$$\min_{\mathbf{D},\alpha_i} \sum_{i=1}^{n} \left(\frac{1}{2} \parallel x_i - \mathbf{D}\alpha_i \parallel^2 + \lambda \parallel \alpha_i \parallel_1 \right), \; s.t., \; \mathbf{D} \geq 0, \; \alpha_i \geq 0. \tag{1}$$

while restricting the atoms to have a norm of less than one. The optimization is achieved through an iterative approach consisting of two alternative steps: the sparse coding step on a fixed $\mathbf{D}$ and the dictionary update step on fixed α_i [13]. As mentioned above, we learn two types of dictionaries: (1) a concept dictionary $\mathbf{D}_c$; (2) an image dictionary $\mathbf{D}_i$.

3.1.4 Dictionary coherence computing

The natural way to measure the degree of coherence C_i between the image dictionary $\mathbf{D}_i$ and the concept dictionary $\mathbf{D}_c$, is to inspect the product matrix: $\mathbf{D}_i^{\mathbf{T}}\mathbf{D}_c$, where the superscript T denotes the matrix transposition. This is because the element d_{ij} of the product matrix represents the inner product between a pair of the two dictionary atoms, i.e., $d_{ij} = d_i \cdot d_j$, here, $d_i \in \mathbf{D}_i$, $d_j \in \mathbf{D}_c$. Therefore, as shown in (2), we compute dictionary coherence C_i through a Frobenius norm defined as the square root of the sum of the absolute squares of the matrix's elements d_{ij}:

$$C_i = \parallel \mathbf{D}_i^{\mathbf{T}}\mathbf{D}_c \parallel_F = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{k} |d_{ij}|^2} \tag{2}$$

where the subscript F denotes the Frobenius norm.

3.1.5 Adaptive sampling

After computing the dictionary coherence C_i between the current web image and the concept, we can easily determine whether to add the current web image to the training set by simply comparing the C_i with a pre-given threshold C_{th}. If $C_i \geq C_{th}$, meaning that the web image is coherent with the concept, then we accept it. Otherwise, we discard it.

Here, we propose an adaptive off-line method through automatic calculation of the threshold C_{th} from the distribution of the coherence degrees of all the positive train samples. According to the theory of hypothesis testing, the threshold C_{th} can be adaptively determined by:

$$C_{th} = \mu - \eta\sigma, \tag{3}$$

where μ and σ are the mean and standard deviation of all the coherence degrees C_{Pos} between each positive training sample and the concept, and η is an empirical parameter that can be determined universally. In our experiments, we set $\eta = \sqrt{3}$.

3.2 Group sparse ensemble learning

3.2.1 Problem formulation

Ensemble learning refers to the process of combining multiple classifiers to provide a single and unified classification decision [25]. Recent research has demonstrated that a good ensemble of localized classifiers can outperform a single (best) classifier learned over the entire dataset [7, 25]. Additionally, learning a set of "smaller" localized classifiers is usually more efficient in terms of algorithmic complexity than a global classifier, which has motivated researchers to adopt the ensemble learning approach for concept detection [25]. [7] advocates to "learn many models not just one".

In visual concept detection, an image or a keyframe of a video shot is processed to detect the presence of a set of predefined concepts. Without loss of generality, we assume that the data instances are represented as vectors, such as the visual feature vectors of keyframes. Mathematically, we denote the observed data matrix as $\mathbf{X} = [x_1, x_2, \ldots, x_n] \in \Re^{d \times n}$, where $x_i \in \Re^d$ represents the i-th data instance vector with dimensionality d. For each concept, we have the label $Y = \{y_i \in \pm 1, i = 1, 2, \ldots, n\}$ for the training set matrix $\mathbf{X}$. Consequently, with the binary classification in the framework of SVM, the ensemble discriminant function $F(x_t)$ for a given test sample x_t is [25]:

$$F(x_t) = \sum_{c=1}^{k} \Psi_c(x_t) \cdot \left(\sum_{i \in \pi_c} \beta_i y_i \langle \Phi(x_t), \Phi(x_i) \rangle + b_c \right) \tag{4}$$

where k localized classifiers are built on instance localities π_c, and $\Psi_c(x_t)$ are the gating functions that govern how localized classifiers are coordinated for the final classification of test sample x_t. Learning the ensemble discriminant function $F(x_t)$, i.e., ensemble construction, can be decomposed into two steps [25]: (1) learning the instance localities π_c and gating function $\Psi_c(x_t)$, and (2) training the individual classifiers to estimate the kernel classifier parameters such as the optimal classification hyperplane parameters β_i, b_c.

3.2.2 Framework of group sparse ensemble learning

Here, we propose to construct the ensemble through AutoGSC [26] to take advantage of the hidden group structures of data. The overall framework is illustrated in Fig.3.

Specifically, as shown in the figure, after feature extraction, we use AutoGSC to learn both a common shared dictionary $\mathbf{D}^S$ over different data groups and the k individual group-specific dictionaries $\left\{\mathbf{D}_c^I\right\}_{c=1}^k$ which can help us to capture the discrimination information contained in the different data groups. We then represent each data instance x_i by using a sparse linear combination of both dictionaries, i.e., get the shared sparse code matrices $\left\{\mathbf{G}_c^S\right\}_{c=1}^k$ from the common shared dictionary $\mathbf{D}^S$ and the individual sparse code matrices $\left\{\mathbf{G}_c^I\right\}_{c=1}^k$ from the individual group-specific dictionaries $\left\{\mathbf{D}_c^I\right\}_{c=1}^k$ for data matrix $\mathbf{X}$. Finally, we compute the reconstruction errors of the group sparse coding for each data instance, and use them to calculate the gating functions $\Psi_c(x_t)$ which are used for instance

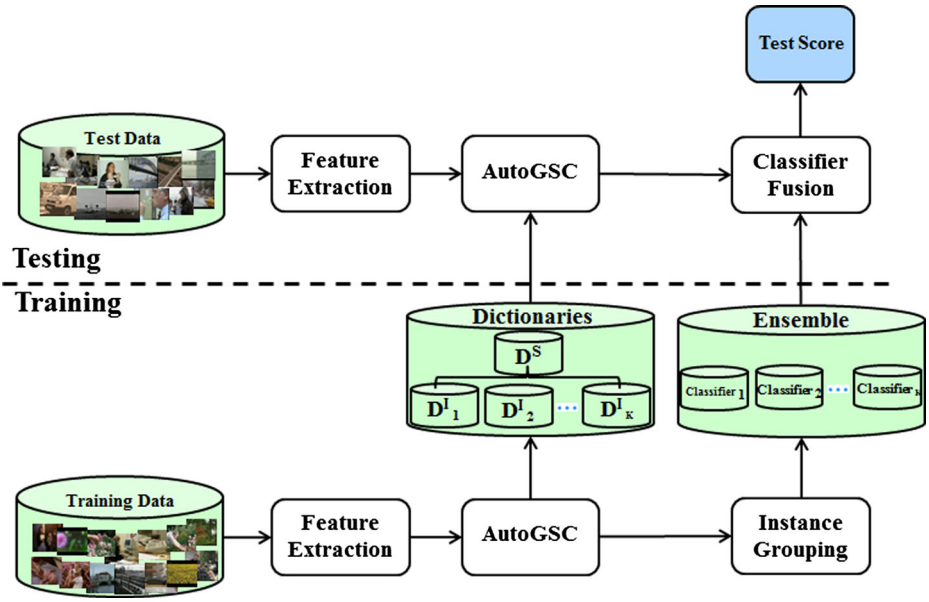

Fig. 3 Proposed Framework of Group Sparse Ensemble Learning

grouping during ensemble construction and individual classification result fusion during testing.

The following subsections will detail the gating function calculation with AutoGSC, ensemble construction and fusion, and complexity analysis.

3.2.3 Gating function calculation with autoGSC

AutoGSC tries to discover the hidden group structures of data by solving the optimization problem under the following non-negativity constraints [26]:

$$\min \sum_{c=1}^{k} \left\| \mathbf{X}_c - \mathbf{D}^S \mathbf{G}_c^S - \mathbf{D}_c^I \mathbf{G}_c^I \right\|_F^2 + \sum_{c=1}^{k} \left[\gamma_I \phi \left(G_c^I \right) + \gamma_S \phi \left(G_c^S \right) \right]$$
$$s.t. \mathbf{D}^S \geq 0, \forall c = 1, 2, \ldots, k, \mathbf{D}_c^I \geq 0, \mathbf{G}_c^I \geq 0 \tag{5}$$

where the normalized matrices $\mathbf{D}^S$ and $\left\{\mathbf{D}_c^I\right\}_{c=1}^k$ to be solved are the common shared dictionary and k individual group-specific dictionaries on each group locality π_c respectively. The matrices $\left\{\mathbf{G}_c^S\right\}_{c=1}^k$ and $\left\{\mathbf{G}_c^I\right\}_{c=1}^k$ to be solved are the sparse code (i.e., reconstruction coefficient) matrices decoded by the two dictionaries correspondingly. In order to achieve group sparsity, the second term is introduced to impose some regularization on the sparse code where function ϕ is used to compute the ℓ_1-norms of the row vectors of the input matrix.

After merging the two kinds of dictionaries and corresponding sparse codes by:

$$\mathbf{D}_c = \left[\mathbf{D}^S, \mathbf{D}_c^I\right]$$
$$\mathbf{G}_c = \left[\mathbf{G}_c^S, \mathbf{G}_c^I\right] \tag{6}$$

the objective function $\hat{\mathcal{J}}_0$ of the optimization problem (5) can be rewritten as:

$$\begin{aligned}\hat{\mathcal{J}}_0 &= \sum_{c=1}^{k}\left[\left\|\mathbf{X}_c - \mathbf{D}_c\mathbf{G}_c^{\top}\right\|_F^2 + \gamma\phi\left(\mathbf{G}_c\right)\right] \\ &= \sum_{c=1}^{k}\sum_{x_i\in\pi_c}\left[\left\|x_i - \sum_j G_{cij}\mathbf{D}_{c\cdot j}\right\|_F^2 + \gamma\sum_j\left|G_{cij}\right|\right]\end{aligned} \tag{7}$$

where $\mathbf{D}_{c\cdot j}$ is the j-th column of $\mathbf{D}_c$, and G_{cij} is the $(i,\ j)$-th entry of $\mathbf{G}_c$.

AutoGSC uses a Lloyd style algorithm [26] to solve to the problem by alternating between the dictionary and sparse code.

Since AutoGSC searches each locality π_c to obtain the group identity of a data instance x_i with the minimum reconstruction error, we propose to use the reconstruction error to calculate the gating function vector $\Psi(x_i) = \{\Psi_c(x_i)\}_{c=1}^{k}$ in the four steps shown in Fig.4. This makes our proposed algorithm very different from the gating function calculation method proposed in [25], which uses the sparse code to obtain the gating function directly.

To achieve high efficiency, especially testing efficiency, we adopt the Instance-Locality Assignment Algorithm proposed in [24, 25] in our proposed algorithm to detect sharp decrease in the two adjacent elements and remove the rest small elements of the descending membership vector. In this paper, we restrict the maximum number of groups to which data instance x_i can be assigned with the input replication parameter l, i.e., an instance can only be allocated to at most l group localities. After calling the algorithm, the returned gating function $\Psi_c(x_i)$ has only r ($r \le l \ll k$) non-zero elements, which means x_i can be allocated into only r groups at the same time.

Algorithm: Gating Function Calculation

Input: Instance x_i, Dictionaries $\{\mathbf{D}_c\}_{c=1}^{k}$, and replication parameter l.

Output: Gating function vector $\Psi(x_i)$.

1: Measure the reconstruction error vector $\{\varepsilon_{ic}\}_{c=1}^{k}$ for quantizing x_i with dictionary $\mathbf{D}_c$ by:

$$\varepsilon_{ic} = \|x_i - \mathbf{D}_c \cdot g_{ci}\|^2 + \gamma\,|g_{ci}|_1 \tag{8}$$

where g_{ci} is the corresponding sparse code of x_i with dictionary $\mathbf{D}_c$, i.e., i-th column of $\mathbf{G}_c$.

2: Compute the grouping weight w_{ic} of assigning x_i to the group locality π_c as the reciprocal of the error, i.e., $w_{ic} = 1/\varepsilon_{ic}$. Thus, we obtain the weight vector $w_i = \{w_{ic}\}_{c=1}^{k}$ over all groups.

3: In order to detect sharp decrease in the two adjacent elements and remove the rest small elements of the descending membership vector, call the Adaptive Instance-Locality Assignment Algorithm proposed in [25] with the input parameter l to get the membership vector $\alpha_i = \{\alpha_{ic}\}_{c=1}^{k}$ from w_i.

4: ℓ_1-normalize membership vector to return the gating function vector $\Psi(x_i) = \{\alpha_{ic}/\,|\alpha_i|_1\}_{c=1}^{k}$.

Fig. 4 Gating Function Calculation Algorithm

3.2.4 Ensemble construction and fusion

For ensemble construction, we use $\Psi_c(x_i)$ to allocate each training data instance x_i into multiple r group localities π_c among all k hidden groups, and use SVM to train the individual classifiers for estimating the kernel classifier parameters such as the optimal classification hyperplane parameters β_i, b_c.

Whereas for ensemble fusion, with (4), we use $\Psi_c(x_t)$ to coordinate the classification results of the related r group localities to obtain the ensemble discriminant function $F(x_t)$ for a given test sample x_t.

3.2.5 Complexity analysis

Similar to the computational complexity analysis in [25], since the theoretical computational complexity of SVM training is between $O\left(n^2\right)$ and $O\left(n^3\right)$ (n is the number of training samples) depending on the value of the hyper-parameter C [25], we can greatly improve the training efficiency after we use AutoGSC by partitioning a given training sample into at most l hidden groups. Specifically, the number of training samples in a given group locality π_c is nl/k on average, and we need to train k individual classifier in the ensemble. Therefore, the complexity is greatly reduced to only l^2/k to l^3/k^2 ($l \ll k$) times that of global classification. While for testing, since we only invoke at most l individual classifiers in the ensemble for a given test sample, and the classifiers are very compact with many fewer support vectors than those used by a global classifier, the testing efficiency can be greatly improved compared with global classification. Thus we conclude that our proposed group sparse ensemble learning method is very efficient in both training and testing.

4 Experiments and results

4.1 Experiment setup

To evaluate the performance of our proposed web image sampling and group sparse ensemble learning methods, we selected the same TRECVid 2008 video benchmark collection [15] as [25] to conduct our experiments. TRECVid is now widely regarded as the actual standard for evaluation the performance of concept based video retrieval systems [25]. The number of positive training samples for each concept in the TRECVid 08 development set is shown in the column "#DPos" of Table 1 [25]. Refer to [15, 25] for more details about the dataset.

4.2 Web image sampling

First, we used the Google API to search and download the top 1000 web images for each concept by constructing a query with the concept name. Then we annotated the images manually; the number of positive samples for each concept in the initial web image set is shown in the column "#WPos" of Table 1. Finally, we used our proposed sampling method to select the positive samples for each concept; the number of positive samples for each concept selected from the web images is shown in the column "#SPos" of Table 1. To test the effectiveness of our proposed method, we performed three runs for each concept:

– **[Baseline]**: Use only positive training samples in the TREC-Vid 08 development set ("#DPos" in Table 1).

Table 1 The number of positive samples for 20 concepts in TRECVid 08

ID	Concept	#DPos	#WPos	#SPos
1001	Classroom	241	790	347
1002	Bridge	186	420	235
1003	Emergency-Vehicle	103	151	11
1004	Dog	136	795	123
1005	Kitchen	289	537	174
1006	Airplane-flying	80	395	113
1007	Two-people	4140	729	458
1008	Bus	106	902	312
1009	Driver	302	489	157
1010	Cityscape	331	879	623
1011	Harbor	217	261	76
1012	Telephone	203	557	412
1013	Street	1799	693	508
1014	Demonstration-Or-Protest	159	68	25
1015	Hand	1879	384	302
1016	Mountain	265	507	284
1017	Nighttime	490	594	229
1018	Boat-Ship	506	783	215
1019	Flower	620	948	513
1020	Singing	441	646	187

Note: The column "#DPos" denotes the number of positive training samples in the TRECVid 08 development set, "#WPos" in the initial positive web image set, "#WPos" in the final web image set after sampling.

- **[AddWeb]**: Use positive training samples of the TREC-Vid 08 development set and the initial positive web image set ("#DPos+#WPos" in Table 1).
- **[Sampling]**: Use positive training samples of the TREC-Vid 08 development set and the web image set after the proposed sampling("#DPos+#SPos" in Table 1).

In the above runs, we used the SIFT features [12] for dictionary learning during sampling, and the well-known BoW feature [10] based on soft-weighting of SIFT, due to its widely reported effectiveness [25].

Figure 5 shows the comparison results of AP for each concept and mean AP (MAP) of the three runs. As shown, the proposed run [Sampling] achieved the highest MAP of 0.144, which is 9.92 % higher than the run [Baseline] (MAP 0.131), and 6.67 % higher than the run [AddWeb] without sampling(MAP 0.135). In particular, the proposed method outperformed the others on 9 out of 20 concepts, including Airplane-flying, Dog, Telephone, Demonstration-Or-Protest, Hand, and Flower, which had been selected with sufficient visually-coherent positive samples, while little was gained with the concepts such as Harbor, Kitchen, Bridge, and Emergency-Vehicle because these concepts on the old documentary TRECVid videos may be too outdated for enough positive web samples to be obtained. On the other hand, the run [AddWeb] achieved only a 3.05 % improvement in MAP compared with the run [Baseline].

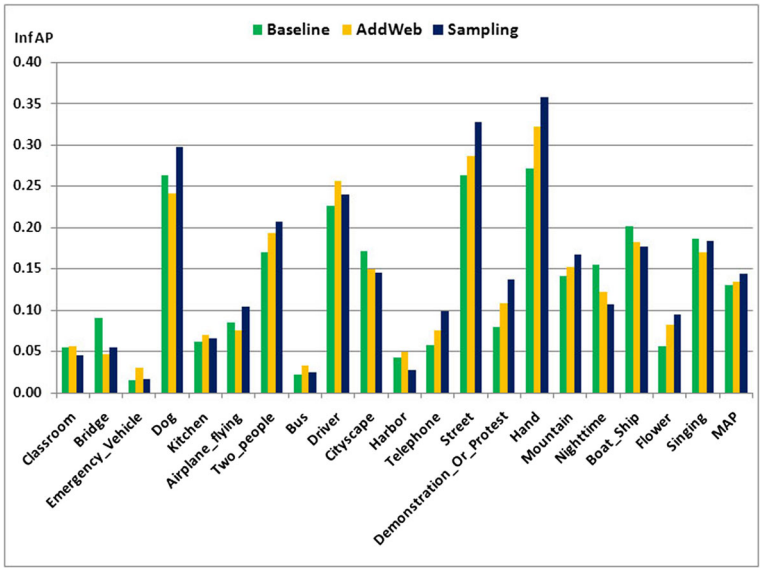

Fig. 5 Comparison results of web image sampling

Compared with the best runs in TRECVid 2008 [10], significant improvement was obtained in handling concepts with few TRECVID positive training samples. The experimental results show that the proposed approach can achieve constant overall improvement despite cross-domain incoherence.

4.3 Group sparse ensemble learning

To test our proposed group sparse ensemble learning method, we conducted comparison experiments with the sparse ensemble learning (SEL) method proposed by [25] we used the same VIREO-374 BoVW features released by [10] as [25] to train and test our system. Additionally, we used the same parameters such as number of locality $k = 800$, replication parameter $l = 20$, RBF kernel of SVM, and same evaluation criteria InfAP as [25] for direct comparison.

First, to verify the effectiveness of our ensemble fusion method based on the proposed gating function calculation algorithm, we compared it with average fusion. InfAP of each concept and MAP yielded by the two fusion schemes are shown in Fig.6. It shows that our fusion method, based on the proposed gating function calculation algorithm, yields MAP of 0.147, 6.5% higher than average fusion (MAP=0.138). We can also see that the proposed method clearly outperforms average fusion on most of the concepts, such as "Dog", "Airplane-flying", "Cityscape", "Street", "Hand", "Mountain", "Nighttime". This superiority does not extend to rare concepts with too few positive training samples such as "Bus" and "Emergency-Vehicle"; their detection rates are low and unstable. Thus we can conclude that our fusion method is effective.

Next, we compared our method with the SEL method [25], global SVM classification (Single-SVM). Ours yielded a MAP of 0.147, an improvement of 4.3 % and 12.2 % over SEL (MAP=0.141) and Single-SVM (MAP=0.131), respectively. Figure 7 compares InfAP

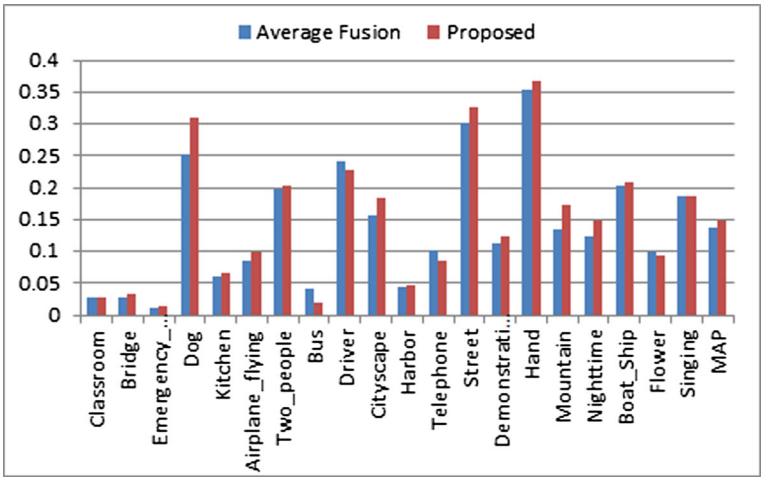

Fig. 6 Fusion comparison with average fusion

for each concept. From the figure, we can see that the proposed method clearly outperforms SEL and single SVM on 11 out of 20 concepts, including both scene concepts like "Cityscape", "Harbor", "Street", "Mountain", "Nighttime" and object concepts like "Dog", "Airplane-flying", "Two-people", "Telephone", "Hand", "Flower". For the remaining concepts such as "driver" and "Demonstration-Or-Protest", SEL performs better, this is due to the diversified patterns of these concepts. Our conjecture is that the instances of these concepts in which our proposed method performs best are not too diversified, and often have good tendency or consistence on being well grouped. This reflects the advantages of group sparse coding in discovering the hidden group structures.

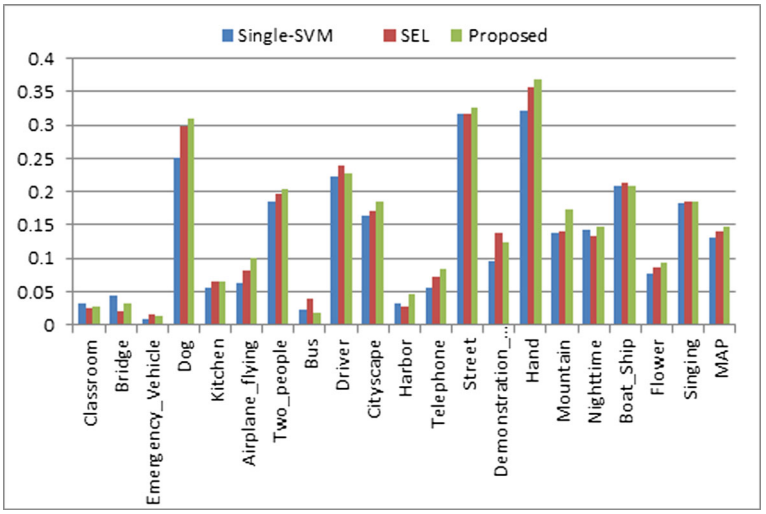

Fig. 7 Comparison with Single-SVM and SEL

Additionally, our experiments also shows that the time complexity of ours is almost equivalent to that of SEL method.

The above results show that the ensemble learning proposal has achieved promising results and can outperform existing approaches.

5 Conclusion

In this paper, we propose a novel web image sampling approach and a novel group sparse ensemble learning approach to tackle the two challenging problems of large scale data collection and training respectively. For data collection, in order to alleviate manual labeling efforts, we propose a web image sampling approach based on dictionary coherence to select coherent positive samples from web images. For efficient training of large scale data, in order to exploit the hidden group structures of data, we propose a novel group sparse ensemble learning approach based on Automatic Group Sparse Coding (AutoGSC). Experiments show that our proposed methods can achieve promising results and outperforms existing approaches.

References

1. Amir A, Berg M, Chang S-F, Hsu W, Iyengar G, Lin C-Y, Naphade M, Natsev AP, Neti C, Nock H, Smith JR, Tseng B, Wu Y, Zhang D IBM research TRECVID-2003 video retrieval system. In: NIST TRECVID Workshop, Nov 2003
2. Bay H, Ess A, Tuytelaars T, Gool LV (2008) SURF: Speeded up robust features. Comp Vision Image Underst 110(3):346–359
3. Bengio DSS, Pereira F, Singer Y (2009) Group Sparse Coding. In: Neural Information Processing Systems - NIPS
4. Bordes A, Ertekin S, Weston J, Bottou L (2005) Fast kernel classifiers with online and active learning. J Mach Learn Res 6:1579–1619
5. Borth D, Ulges A, Breuel TM (2011) Automatic concept-to-query mapping for web-based concept detector training. In: ACM Multimedia 2011, pp 1453–1456
6. Cao J, Lan Y, Li J, Li Q, Li X, Lin F, Liu X, Luo L, Peng W, Wang D, Wang H, Wang Z, Xiang Z, Yuan J, Zhang B, Zhang J, Zhang L, Zhang X, Zheng W Intelligent multimedia group of Tsinghua University at TRECVID, 2006. In: NIST TRECVID Workshop, Nov 2006
7. Domingos P (2012) A few useful things to know about machine learning. Commun ACM 55(10):78–87
8. Enzweiler M, Gavrila DM (2009) Monocular pedestrian detection: Survey and experiments. IEEE Trans Pattern Anal Mach Intell 31:2179–2195
9. Huiskes MJ, Thomee B, Lew MS (2010) New trends and ideas in visual concept detection: the MIR Flickr retrieval evaluation initiative. In: Proceedings of the international conference on Multimedia Information Retrieval (MIR 2010), pp 527–536
10. Jiang Y-G, Yang J, Ngo C-W, Hauptmann AG (2010) Representations of keypoint-based semantic concept detection: A comprehensive study. IEEE Trans Multimed 12(1):42–53
11. Li H, Wang X, Tang J, Zhao C (2013) Combining global and local matching of multiple features for precise retrieval of item images. ACM/Springer Multimed Syst J 19(1):37–49
12. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
13. Mairal J, Bach F, Ponce J, Sapiro G (2010) Online learning for matrix factorization and sparse coding. J Mach Learn Res 11:19–60
14. Munder S, Gavrila D (2006) An experimental study on pedestrian classification. IEEE Trans Pattern Anal Mach Intell 28:1863–1868
15. Over P, Awad G, Rose RT, Fiscus JG, Kraaij W, Smeaton AF (2008) Trecvid 2008 - goals, tasks, data, evaluation mechanisms and metrics. In: NIST TRECVID Workshop
16. Pytlik B, Ghoshal A, Karakos D, Khudanpur S TRECVID 2005 Experiment at Johns Hopkins University: Using Hidden Markov Models for Video Retrieval. In: NIST TRECVID Workshop, Nov 2005

17. Ramirez I, Sprechmann P, Sapiro G (2010) Classification and clustering via dictionary learning with structured incoherence and shared features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), pp 3501–3508
18. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
19. Song Y, Zheng Y-T, Tang S, Zhou X, Zhang Y, Lin S, Chua T-S (2011) Localized multiple kernel learning for realistic human action recognition in videos. IEEE Trans Circ Syst Vi Technol 21(9):1193–1202
20. Sun Y, Kojima A (2011) A novel method for semantic video concept learning using web images. In: ACM Multimedia 2011, pp 1081–1084
21. Sun Y, Shimada S, Taniguchi Y, Kojima A (2008) A novel region-based approach to visual concept modeling using web images. In: ACM Multimedia 2008, pp 635–638
22. Tang S, Li J-T, Li M, Xie C, Liu Y, Tao K, Xu S-X Trecvid 2008 high-level feature extraction by MCG-ICT-CAS. In: NIST TRECVID Workshop, Nov 2008
23. Tang S, Li J-T, Zhang Y-D, Xie C, Li M, Liu Y, Hua X, Zheng Y-T, Tang J, Chua T-S PornProbe: an LDA-SVM based pornography detection system. In: ACM Multimedia 2009, Oct. 2009
24. Tang S, Zheng Y-T, Cao G, Zhang Y-D, Li J-T (2012) Ensemble learning with LDA topic models for visual concept detection. Multimedia - A Multidisciplinary Approach to Complex Issues, pp 175–200
25. Tang S, Zheng Y-T, Wang Y, Chua T-S (2012) Sparse ensemble learning for concept detection. IEEE Trans Multimed 14(1):43–54
26. Wang F, Lee N, Sun J, Hu J, Ebadollahi S Automatic group sparse coding. In: Twenty-Fifth AAAI Conference on Artificial Intelligence, Aug 2011
27. Zha Z-J, Wang M, Zheng Y-T, Yang Y, Hong R, Chua T-S (2012) Interactive video indexing with statistical active learning. IEEE Trans Multimed 14(1):17–27
28. Zha Z-J, Zhang H, Wang M, Luan H, Chua T-S (2013) Detecting group activities with multi-camera context. IEEE Transactions on Circ Syst Vi Technol 23(5):856–869
29. Zhu S, Wang G, Ngo C-W, Jiang Y-G (2010) On the sampling of web images for learning visual concept classifiers. In: Proceedings of the 9th ACM International Conference on Image and Video Retrieval (CIVR 2010), pages 50–57, Xi'an, China

Yongqing Sun received the B.E. and M.E. degrees in Computer Science from Xi'an Jiaotong University (XJTU), China in 1998 and 2001 respectively, and the Ph.D. degree in Information and Computer Science from Keio University, Japan in 2005. She is currently a researcher in NTT Media Intelligence Laboratories. Her research interests include image processing, pattern recognition, machine learning, data mining and multimedia retrieval.

Kyoko Sudo is now a senior research engineer in NTT Media Intelligence Laboratories. She received the B.E. and M.E. degrees in mathematical engineering and information physics and a Ph.D. in informaiton physics and computing from Tokyo Univerisity, Tokyo, in 1991,1993, and 2007, respectively. Since joining NTT Laboratories in 1993, she has been engaged in research on image processing and pattern recognition.

Yukinobu Taniguchi received the B.E., M.E., and Dr.Eng. degrees in mathematical engineering from the University of Tokyo in 1990, 1992, and 2002, respectively. He joined NTT Corporation in 1992 He is currently a senior research engineer, supervisor, of NTT Media Intelligence Laboratories. His research interests include image/video processing and multimedia applications.