Contextual Decomposition of Multi-Label Images

Teng Li [1], Tao Mei [2], Shuicheng Yan [3], In-So Kweon [1], and Chilwoo Lee [4]
[1] Dept. of Electrical Engineering, KAIST, Daejeon, Korea
[2] Microsoft Research Asia, Beijing, P. R. China
[3] Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore
[4] Dept. of Electronics and Computer Eng., Chonnam Nat'l Univ., Gwangju, Korea
tengli@rcv.kaist.ac.kr, tmei@microsoft.com, eleyans@nus.edu.sg,iskweon@ee.kaist.ac.kr, leecw@chonnam.ac.kr

Abstract

Most research on image decomposition, e.g. *image segmentation and image parsing, has predominantly focused on the low-level visual clues within single image and neglected the contextual information across different images. In this paper, we present a new perspective to image decomposition piloted by the multi-labels associated with individual images. Observing that the context information (*i.e.*, local* label representations *of the same label are similar while those from different labels are dissimilar) exists across different images, we propose to perform image decomposition in a collective way, and then the image decomposition problem is formulated as an optimization which maximizes inter-label difference and at the same time minimizes intra-label difference of the target label representations. Such contextual image decomposition has a wide variety of applications, among which the two exemplary ones are: 1) multi-label image annotation in which the sparse coding of a query image over the bases consisting of all learned label representations naturally produces the multi-label annotation, and 2) label ranking in which the annotated labels are re-ordered according to the sparse coding coefficients on those learned label representations. It is worth noting that these two applications can be performed simultaneously via the label propagation process in sparse coding.*

1. Introduction

Image decomposition, which parses an image into a set of constitutive components (*e.g.*, objects, parts, surfaces, and primitives), is a fundamental and important step for solving many high-level vision problems, such as scene understanding, object detection and recognition, and image synthesis.

Most conventional approaches to image decomposition focused on low-level visual cues, such as a bottom-up image segmentation in which pixels are locally grouped on the basis of their appearances [14] and a top-down image parsing in which primitives (e.g., rectangles, sketches, and edges) are correlated based on a few grammar rules [15]. Moreover, most of them treat each image as an independent entity without considering the context information across different images.

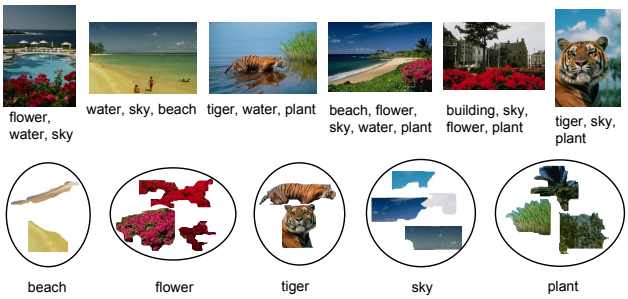

Figure 1. Observations on real-world images (from Corel dataset [18]). The first row shows the original images and their labels. The second row shows the visual representations of different classes/labels, *i.e.*, label representations. It is observed that label representations of the same class are similar while those from different classes are dissimilar, which is what we refer to as "context" in this paper. The optimal decomposition of images in terms of high-level semantic should take both the multi-label setting and context information into account.

In this work, we present a novel approach to image decomposition task driven by high-level contextual information across different images. The proposed approach is motivated by the following observations on real-world images.

1. An image is typically associated with multiple labels, and its visual representation reflects the combination of the involved labels. Figure 1 shows some images and their corresponding labels. It is observed that each label corresponds to certain local patch in the image. Therefore, the entire *image representation* can be decomposed as a set of local *label representations* corre-

sponding to the labels associated with an image.

2. Contextual information is embedded across different images. From Figure 1, we can see in the second row that the local representations of the same label, which are grouped, depict similar visual appearance, while those from different labels show to be different from each other. Therefore, the decomposition of image can be guided by this contextual information on the basis of a sufficient collection of images.

In the proposed approach, the image representation is decomposed to local label representations automatically *without explicit segmentation* in a collective way [1] [2]. The learned label representations should be representative, discriminative, and comprehensive so that the original image can be semantically recovered. As shown in Figure 2, an image representation is a linear combination of local label representations, each weighted by the presence of the corresponding label associated to this image. For example, the first image can be decomposed into the linear combination of the label representations of "cow," "sky," and "grass." To this end, we learn an optimal set of label representations from a collection of training images by considering the context information across images, and formulate it into an optimization problem by minimizing the intra-label difference while maximizing the inter-label difference of all the label representations.

The derived label representations can be used in many high-level vision tasks. Figure 2 shows two exemplary applications: multi-label image annotation and label ranking. In the first application, multi-label image annotation can be achieved based on the sparse representation classification (SRC) theory which estimates the labels of a testing image by considering it as sparse coding, derived by ℓ^1-reconstruction [16], of all the learned label representations from the training images. In the other application, the sparse coding coefficients naturally indicate the relevance of the labels to the test image. We have evaluated these applications on two widely adopted image benchmarks, Microsoft Research Cambridge (MSRC) and Corel image datasets.

It is worth noting that our approach to multi-label image decomposition is different from conventional image segmentation and parsing in that the proposed contextual image decomposition is driven by high-level visual problems such as image annotation rather than pixel-based primitive

[1] We denote the local representation which corresponds to one of the labels associated to the image as "label representation" to differentiate "image representation" which indicates the visual representation of the whole image.

[2] The visual representation comes into a variety of forms. For example, the Bag-of-Words (BoW) is an effective representation by considering the distribution of local features [20], the color histogram, and so on. In this paper, we adopt the BoW for visual representation owing to its robustness to scale and rotation.

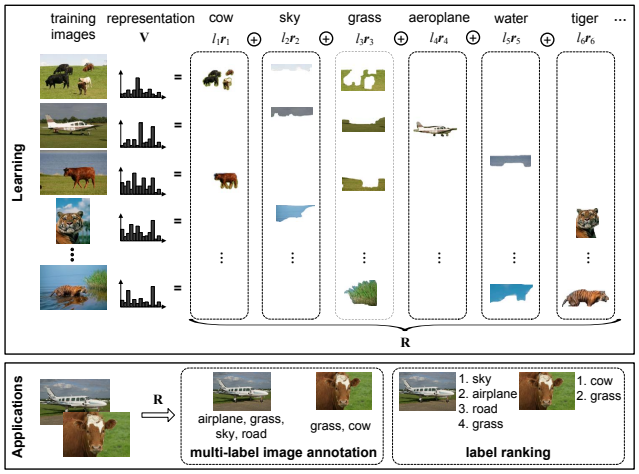

Figure 2. Illustration of the main idea of contextual image decomposition. The learning stage decomposes the image representations into a set of local label representations. Given the learned label representations, we have a variety of applications based on the sparse coding of SRC, such as multi-label image annotation and label ranking. It is worth noting that no segmentation is required in this proposed approach.

grouping [14] [15]. Furthermore, we leverage the context information which is a common observation on real-world images but neglected in conventional approaches. From this point view, it is a kind of collaborative approach to decompose images. From the application perspective, our approach to multi-label image annotation is essentially different from the existing works in which a time-consuming and error-prone segmentation step is required [7] [17] [19]. Moreover, we can achieve label ranking which reords the labels in terms of importance or relevance to image content, and cannot be well obtained by the state-of-the-art annotation approaches and has not been deeply studied yet.

2. Related Works

The related works to this paper include those for traditional image decomposition such as image segmentation and parsing, as well as the applications based on image decomposition such as multi-label image annotation and label ranking.

2.1. Image Segmentation and Parsing

In a broad view, image decomposition can be regarded as transforming a two-dimensional signal using a sequence (cascade) of bottom-up filters. Traditional image decomposition includes image segmentation which solves the perceptual grouping of pixels based on the local appearances [14] and image parsing which poses an image into its constituent local patterns in a hierarchical tree [15]. However,

most of them have focused on low-level vision tasks. In other words, the filters in previous approaches are designed in a bottom-up manner on the basis of low-level appearance. Moreover, they treat each image independently without considering the semantic correlation or context across multiple images. As a result, the bottom-up decomposition approaches achieved limited success, especially in high-level vision applications.

The proposed contextual approach in this paper is a new perspective to image decomposition which is driven by understanding the real-world images with multiple labels, and collaboratively leveraging the context information across different images.

2.2. Multi-Label Image Annotation

Recently, with the popularity of photo sharing sites like Flickr [1], multi-label image annotation is becoming increasingly important owing to the great potential in automatically tagging images with related labels. Conventional image classification methods usually consider an image as an entity associated with only one label in model learning stage [3]. One can do multi-label classification using this kind of methods by running the binary classifier on the images for each label one by one. However, a real-world image usually contains multiple classes rather than a single one. The image representations are therefore not effective to capture the multi-label characteristics, which in turn defects the performance of multi-label classification.

Methods modeling the relation between labels and regions were also proposed, usually based on the multi-instance learning (MIL) framework [11]. An image is regarded as a bag and then divided to sub-images or segments. With labels at the image level rather than region level, MIL finds which sub-images contribute to which labels of the image. Conventional MIL approaches mainly focus on the single label problems. To solve the multi-instance multi-label problem, methods were proposed to simultaneously model the relations between the labels and regions, as well as the correlations among the multiple labels [19, 21]. In these MIL methods, image segmentation is an indispensable step, which is still error-prone and significantly augments the computational cost. Although recent works have attempted to simultaneously perform image segmentation and annotation [4, 12], a preprocess step of over-segmentation is still critical in these methods. The application to multi-label image annotation on the basis of the decomposed label representations in this paper does not require a segmentation step. Instead, the proposed data decomposition approach can be regarded as a kind of implicit segmentation process.

2.3. Label Ranking

Most existing photo sharing sites only contain a list of orderless tags associated with each image. The different importance or relevance degrees of these tags are neglected, which limits the effectiveness of these labels in image search and other applications. Although a recent work has proposed to mine the ranking information based on correlative visual information [10], it is assumed that the labels of each image are given. Simultaneous image annotation and label ranking still remain a challenge problem.

3. The Approach

3.1. Feature Representation

In this work, the images and regions are represented using the BoW [20]. A set of local features are extracted from each image or region. Then a visual vocabulary is learned by k-means clustering over the training features, where each cluster centroid is set as a visual word. Assigning each feature to its nearest visual word and counting the occurrence number, we can represent an image by a histogram of visual words. The number of visual words is set as 500 in the experiments for the tradeoff of classification accuracy and computational cost. Local patches centering on a regular grid with spacing 8 pixels are extracted as the features and their descriptors are computed. The texton histogram descriptor [9] encoding texture and color information is used to represent the local features. Pixel features from the response of several filter banks are clustered to textons and the normalized texton histogram within a segmented local region is computed as the descriptor. In the experiments, 200 textons are used and the descriptor therefore is with 200 dimensions.

3.2. Problem Formulation

Given N training images with multi-labels, we aim to derive the representations for the labels from the image representations. Let $n \in \{1, \ldots, N\}$ denote the training images and $c \in \{1, \ldots, C\}$ denote the label classes. For a specific image n, its representation can be calculated as a K-dimension normalized vector $\mathbf{v}^n \in \mathcal{R}^K$ with $||\mathbf{v}^n||_1 = 1$, where K is the dimension of the BoW. The labels of this image are denoted in a vector $\mathbf{l}^n \in \mathcal{R}^C$ with only 0 or 1 elements:

$$\mathbf{l}_c^n = \begin{cases} 1 & \text{if label } c \text{ is contained in image } n; \\ 0 & \text{if there is no label } c. \end{cases}$$

Let a K-dimension vector $\mathbf{r}_c^n \in \mathcal{R}^K$ denote the representation of the corresponding part for the c-th label in image n. The vectors $\{\mathbf{r}\}$ for representing all the non-zero labels in the N images are the objective to be computed.

The image representation $\mathbf{v}^n$ can be written as the linear combination of the label representations in the image as

shown in Figure 2, which is formulated as follows:

$$\forall n, \begin{cases} \mathbf{v}^n = \sum_{c=1,\mathbf{l}_c^n=1}^{C} \mathbf{r}_c^n. \\ \mathbf{r}_c^n(k) \geqslant 0, k \in \{1,...,K\}, c \in \{1,...,C\}. \end{cases} \quad (1)$$

Note that $\mathbf{r}_c^n$s are not normalized and $||\mathbf{r}_c^n||_1 \in [0,1]$. $||\mathbf{r}_c^n||_1$ can be considered as the portion of the region of label c in image n. Eq. (1) encloses the information on training set, *i.e.*, different combinations of label representation result in different observations of images.

Since label representations of the same label should be similar to each other while those of different labels should be dissimilar, the optimal vectors $\{\mathbf{r}\}$ are obtained by minimizing the objective function defined in terms of minimizing intra-label difference while maximizing inter-label difference as follows:

$$\sum_{n=1}^{N}\sum_{c=1,\mathbf{l}_c^n=1}^{C}\left\{\sum_{m=1,\mathbf{l}_c^m=1}^{N}\frac{||\frac{\mathbf{r}_c^n}{|\mathbf{r}_c^n|}-\frac{\mathbf{r}_c^m}{|\mathbf{r}_c^m|}||_2^2}{M_{nc}} -\eta\sum_{m=1}^{N}\sum_{c'=1,c'\neq c,\mathbf{l}_{c'}^m=1}^{C}\frac{||\frac{\mathbf{r}_c^n}{|\mathbf{r}_c^n|}-\frac{\mathbf{r}_{c'}^m}{|\mathbf{r}_{c'}^m|}||_2^2}{M'_{nc}}\right\}, \quad (2)$$

where the first part measures the intra-label difference and the second part measures the inter-label difference for all the label representations of the training images, and η is a weighting factor. M_{nc} and M'_{nc} represent the number of samples from class c and the number of samples of other classes. Considering Eq. (2) and Eq. (1) together, the problem is formulized as an constrainted optimization problem where Eq. (1) serves as the constraints. In the next we will give the solution to Eq. (1).

3.3. Learning to Decompose

To solve the optimization problem defined by Eq. (2) with the constraints from Eq. (1) directly is computationally intractable for large training set, since each image introduces a set of constraints in Eq. (1) and additional computation costs for Eq. (2). Here we adopt an iterative procedure described in Algorithm 1 which optimizes the objective function with respect to one image each time.

As in step 3 of Algorithm 1, for the current processing image n, assuming the decomposed label representations from other images, i.e. $\{\mathbf{r}_c^m\}, m \in \{1,\ldots,N\}, m \neq n$, are known. the optimal $\{\mathbf{r}_c^n\}$ for this image is obtained by minimizing the following function:

$$F(\{\mathbf{r}_c^n\}) = \sum_{c=1}^{C}\left\{\sum_{m=1,\mathbf{l}_c^m=1}^{N}\frac{||\frac{\mathbf{r}_c^n}{|\mathbf{r}_c^n|}-\frac{\mathbf{r}_c^m}{|\mathbf{r}_c^m|}||_2^2}{M_{nc}} -\eta\sum_{m=1}^{N}\sum_{c'=1,c'\neq c,\mathbf{l}_{c'}^m=1}^{C}\frac{||\frac{\mathbf{r}_c^n}{|\mathbf{r}_c^n|}-\frac{\mathbf{r}_{c'}^m}{|\mathbf{r}_{c'}^m|}||_2^2}{M'_{nc}}\right\}. \quad (3)$$

Algorithm 1 Contextual image decomposition algorithm

1: Input N training image representations with labels $\{\mathbf{v}^n, \mathbf{l}^n\}_{n=1}^N$. Initialize representations for all the labels $\{\mathbf{r}_c^n\}$, $n \in \{1,\ldots,N\}$, as $\mathbf{r}_c^n = \frac{\mathbf{v}^n}{C'}$, where C' is the number of nonzero elements in $\mathbf{l}^n$.
2: For each image n one by one, fixing $\{\mathbf{r}_c^m\}$ of other images, $m \in \{1,\ldots,N\}, m \neq n$, calculate optimal $\{\mathbf{r}_c^n\}$ by Algorithm 2 and update the label representations. Sum up the decrease of the objective function for all the images.
3: If the decrease of objective function is smaller than a threshold, go to step 3. Otherwise go to step 2.
4: Output the decomposed label representations from the N images $\{\mathbf{r}_c^n\}$, $n \in \{1,\ldots,N\}$.

Algorithm 2 Datum-wise iterative optimization procedure

1: Input the temporary label representations of the n-th image $\{\mathbf{r}_c^n\}$ and other images $\{\mathbf{r}_c^m\}$, $m \in \{1,\ldots,N\}, m \neq n$. Set the iteration counter $k = 1$.
2: Update $\{\mathbf{r}_c^n\}$ by $\{\mathbf{r}_c^n\}^{(k+1)} = \{\mathbf{r}_c^n\}^{(k)} - \lambda \cdot \frac{\partial \psi}{\partial \{\mathbf{r}_c^n\}}$, where λ is the step length of the gradient descent algorithm. Here is a search process for a proper λ:
   2.1 initialize a λ;
   2.2 calculate $\{\mathbf{r}_c^n\}^{(k+1)}$;
   2.3 check whether the objective function ψ decrease, if not, set $\lambda = \lambda/2$ and go to step 2.1.
   If the penalty item is smaller than a threshold at current $\{\mathbf{r}_c^n\}^{(k+1)}$ while the decrease of objective function is smaller than a threshold, go to step 3; otherwise set $k = k+1$ and go to step 2.
3: Output the result $\{\mathbf{r}_c^n\}$.

It is a nonlinear constraint optimization problem to minimize F while satisfying the constraints in Eq. (1). It can be solved by an exterior penalty function method, which converts the problem to an equal non-constraint form by including the constraints as penalty components [2]. As a result, we can get a non-constraint form $\psi(\{\mathbf{r}_c^n\})$ to minimize:

$$\psi(\{\mathbf{r}_c^n\}) = F(\{\mathbf{r}_c^n\}) + \epsilon \cdot \left\{||\mathbf{v}^n - \sum_{c=1,l_c^n=1}^{C}\mathbf{r}_c^n||_2^2 + \sum_{c=1,l_c^n=1}^{C}\sum_{k=1}^{K}[\min\left(0,\mathbf{r}_c(k)\right)]^2\right\}. \quad (4)$$

It can be seen that the penalty components equal to the minimum 0 when $\{\mathbf{r}_c^n\}$ satisfies the constraints, therefore the optimal $\{\mathbf{r}_c^n\}$ for Eq. (3) is also the optimal solution to ψ. ϵ is a penalty factor which forces the solutions to satisfy the constraints. If ϵ is set too small, the constraint func-

tion cannot influence the result effectively. If it is set too large, it causes the difficulties to the minimization of the penalty component. Therefore, usually a series of values $\{\epsilon_k\}$ from small to large are adopted in turn in an iterative optimization process. Optimization by exterior penalty function method is taken in the whole $\mathcal{R}^K$ space and the initialization is not limited. Theoretically the global optimal solution can be obtained with proper settings [2]. The gradient descent method is adopted to minimize the objective function of Eq. (4). It is an iterative procedure described in Algorithm 2. The proposed method can minimize the objective value while satisfying the constraints.

4. Applications

The proposed contextual image decomposition algorithm can be applied in many applications, *e.g*. multi-label image annotation and label ranking. Given the decomposed label representations of the N training images, *i.e*., $\{\mathbf{r}_c^n\}, n \in \{1, \ldots, N\}$, and a test image with image representation $\mathbf{v}^y$, we introduce how to predict its associated labels, *i.e*., nonzero elements of its label vector $\mathbf{l}^y$, and at the same time provide the ranking order of these labels based on the recently proposed SRC method [16].

The SRC assumes that images from a single class lie in a linear subspace, which is consistent with our proposed contextual image decomposition in which image representation is the linear combination of label representations. Here we arrange the normalized label representations as column vectors of the matrix $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_{\tilde{N}}] \in \mathcal{R}^{K \times \tilde{N}}$, and let $\mathbf{R}_c \in \mathcal{R}^{K \times \tilde{N}_c}$ denote the submatrix of $\tilde{N}_c$ label representations of the c-th class/label. Given sufficient label representations decomposed from the training images, a test image $\mathbf{v}^y$ of the c-th class will approximately lie in the linearly spanned space of the label representation from the same class: $\mathbf{v}^y = \mathbf{R}_c\mathbf{x}_c$, where $\mathbf{x}_c \in \mathcal{R}^{\tilde{N}_c}$ is a vector of coefficients. If express $\mathbf{v}^y$ in terms of the entire training set $\mathbf{R}$:

$$\mathbf{v}_y = \mathbf{R}\mathbf{x}, \tag{5}$$

then $\mathbf{x} = [0 \ldots \mathbf{x}_c^T \ldots 0]^T \in \mathcal{R}^{\tilde{N}}$ is a vector with only entries associated with the corresponding class c as nonzero. Seeking for a sparse representation, [16] solves the problem by minimizing the ℓ^1 norm of the vector $\mathbf{x}$ and the residual of linear approximation $\mathbf{e} = \mathbf{v}^y - \mathbf{R}\mathbf{x}$. If $\mathbf{v}^y = \mathbf{R}\mathbf{x} + \mathbf{e}$, and

$$\mathbf{v}^y = [\mathbf{R}\ \mathbf{I}]\begin{bmatrix}\mathbf{x}\\ \mathbf{e}\end{bmatrix} = \mathbf{B}\mathbf{w}, \tag{6}$$

then the ℓ^1 minimization solution is obtained by

$$\widehat{\mathbf{w}}_1 = \arg\min ||\mathbf{w}||_1,\ s.t.\ \mathbf{B}\mathbf{w} = \mathbf{v}^y. \tag{7}$$

Denote the label vector of the test image as $\mathbf{l}^y$, then based on the linear approximation of SRC, the resulted coefficients vector $\mathbf{x}$ should ideally have zero entries except for those corresponding to the right labels. The classification therefore can be performed by

$$\mathbf{l}^y(c) = \begin{cases} 1, & \text{if } ||\mathbf{x}_c||_1 > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The resulted coefficients associated with a class reflects how much the image can be linearly approximated by samples of this class. In the proposed contextual image decomposition algorithm, the $\mathbf{r}_c^n$ in section 3 is normalized, we can evaluate the relevance of label c by:

$$P(c) = ||\mathbf{x}_c||_1. \tag{8}$$

The rank of multiple labels is obtained simultaneously with the image annotation according to this relevance. The proposed label ranking can be taken not only for test images, but also for the training images based on $||\mathbf{x}_c||_1$.

The SRC method has several advantages for these two tasks comparing to conventional classifiers such as Support Vector Machine (SVM): 1) multi-label inference and ranking are taken simultaneously while SVM needs to conduct multiple binary classifications, which causes the computational complexity to increase when the number of classes increases; 2) SRC does not need training while SVM classifiers have to be re-trained if there are new training samples or new classes; and 3) SRC can predict non-zero labels naturally with the sparse representation while SVM gives probabilities for all the classes.

5. Experiments

5.1. Experimental Setup

The first dataset used in the experiments is the MSRC image database [13] which contains 591 images from 23 classes. Around 80% images are associated with more than one label and there are around three labels per image on average. However, some classes have only few positive samples, and some are only coarsely associated with one or few labels. We select a subset of the database, focusing on relatively well labeled classes, yielding 355 images and 14 different classes: "building," "grass," "tree," "cow," "sheep," "sky," "mountain," "airplane," "water," "car," "bicycle," "bird," "road," and "boat." The MSRC dataset also provides pixel level ground truth. We randomly split the dataset into two equal subsets for training and testing with the consideration that each class has enough samples in the training set. The second dataset is a subset of the labeled Corel database used in [18]. This set contains 674 labeled images with 11 labels: "Sky," "Waterscape," "Mountain," "Grass," "Tree," "Flower," "Rock," "Earth,"

"Ground," "Building_Material," and "Animal_Skin". This dataset is also split into the training and test subsets in the same way as in the MSRC dataset. These two datasets do not provide the label rank information. To evaluate the performance in label ranking, we manually rank the labels for each test image as the ground truth according to the region size and conceptual importance. Area under ROC curve (AUC) is adopted to measure the performance of multi-label image annotation.

Previous works have not addressed the problem of image label ranking. To measure the performance, we define a weighted normalized discount cumulative gain (WNDCG) based on the normalized discount cumulative gain (NDCG) [8], which is widely used in evaluating information retrieval systems. Given some relevant keywords and a list of ranked retrieval results of length L, the discount cumulated gain (DCG) is $\sum_{i=1}^{L} \omega(i)\ell(i)$, where $\ell(i) = 1$ if the i-th retrieval result is relevant, and $\ell(i) = 0$ otherwise. $\omega(i)$ is a weighting factor assigning top ranked results the highest weights. $\omega(i) = 1$ when $i < b$, and $\omega(i) = \log(b)/\log(k)$ otherwise where b is a constant number (here $b = 2$). Denoting the DCG value of the best resultant list as DCG_best for normalization, NDCG is obtained by

$$NDCG = \frac{1}{DCG\_best} \sum_{i=1}^{L} \omega(i)\ell(i). \tag{9}$$

In this work, considering the labels of an image have different relevance degrees, we define the WNDCG by assigning them different weights $\tilde{\ell}(i)$. Given an image with a list of ranked labels of length $\tilde{L}$, $\tilde{\ell}(i) = \tilde{L} - r + 1$ if the i-th retrieval label is ranked as the r-th relevant in the ground truth, and $\tilde{\ell}(i) = 0$ if the i-th result is not relevant. The WNDCG is then obtained by

$$WNDCG = \frac{1}{WDCG\_best} \sum_{i=1}^{L} \omega(i)\tilde{\ell}(i). \tag{10}$$

In the experiments, three representations are compared: 1) Im_Rep: BoW of the image as in [20], for each label, all the training images associated with this label are collected as the training samples; 2) Reg_Rep: labeled region representation, since the MSRC dataset provides region-level labels, we crop the labeled regions out from the training images as the training samples; and 3) DC_Rep: our proposed contextual image decomposition, we use the decomposed label representations from the training images as training samples. Note that the Reg_Rep is not applied to the Corel dataset, since this dataset does not provide the region-level labels.

Besides the adopted SRC, the SVM classifier is also applied in experiments to extensively study the effectiveness of the decomposed representations. As aforementioned, SVM has several disadvantages in multi-label image annotation and label ranking applications, it however offered the state-of-the-art classification performances in many tasks. The libSVM library [5] is used and linear kernel is adopted. Probabilities for the multiple classes can be yielded for AUC calculation and label ranking. Moreover, the recent proposed Multi-Label Multi-Instance Learning (MLMIL) [19] algorithm for multi-label classification is also compared here. For MLMIL, the same setting as [19] was used.

5.2. Multi-label Image Annotation Results

Table 1 lists the image annotation performance of different algorithms on the MSRC dataset. Furthermore, each test image is then segmented to 20-30 regions and each region is associated with a single label according to the pixel-level ground truth. In Table 2, we evaluate the performances of these algorithms for region annotation. The comparison of image annotation performances of different algorithms on the Corel dataset is given in Table 3.

The proposed DC_Rep shows to be the best in image annotation tasks, and superior to the conventional Im_Rep in all the cases. In the region annotation task the Reg_Rep yields the highest performance, but note that it requires the region-level supervision. With SVM, the Im_Rep shows high performance for image annotation while the Reg_Rep shows to be poor. Since in the MSRC dataset the background scenes are similar for a class, while the pure regions of some classes have high variation, such as "bird" and "boat". The Im_Rep+SVM utilizes the background information in classification while the Reg_Rep does not. The proposed DC_Rep achieves the best performance since it realizes a reasonable trade-off between Im_Rep and Reg_Rep, *i.e.*, certain noise in Im_Rep is excluded while the intra-label variation in Reg_Rep is decreased. Figure 3 illustrates the detailed results for individual labels of the MSRC dataset. Comparing to the MSRC dataset, images in the Corel dataset have various background, while our proposed method can learn representations more representative than Im_Rep, it yields better performance. Also the completeness of labels for the images in this dataset makes our proposed method more suitable, i.e, Eq. (1) can be well satisfied. Figure 4 illustrates the detailed results for individual labels of the Corel dataset.

The SRC obtains similar AUCs as SVM when using DC_Rep for the theoretic coincidence of linear combination. It also cooperates well with the ground truth Reg_Rep, which validates our assumption. The Im_Rep+SRC shows very poor performance when each image is considered as a whole in training, since in this case there is some overlap between the linear subspaces of samples from different classes. From these results, we can observe that SVM is generally better than SRC in annotation performance. However, since SRC is influenced much by the number of train-

Table 1. The image annotation average AUCs on the MSRC dataset of different methods. Note that MLMIL yields 0.868.

Classifier	Im_Rep	Reg_Rep	DC_Rep
SVM	0.901	0.791	**0.910**
SRC	0.417	0.824	**0.883**

Table 2. The region annotation average AUCs on the MSRC dataset of different methods. Note that MLMIL yields 0.827.

Classifier	Im_Rep	Reg_Rep	DC_Rep
SVM	0.793	**0.911**	0.832
SRC	0.439	**0.875**	0.802

Table 3. The image annotation average AUCs on the Corel dataset of different methods. Note that MLMIL yields 0.699.

Classifier	Im_Rep	DC_Rep
SVM	0.795	**0.831**
SRC	0.523	**0.828**

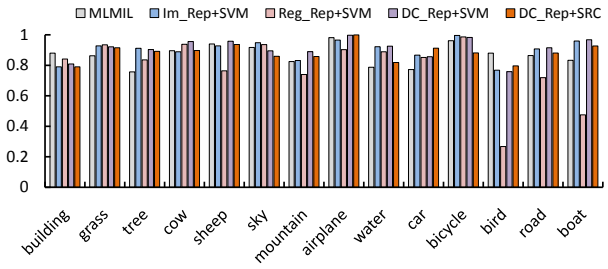

Figure 3. The AUCs of 14 classes of the MSRC dataset using MLMIL, Im_Rep, Reg_Rep and our proposed DC_Rep.

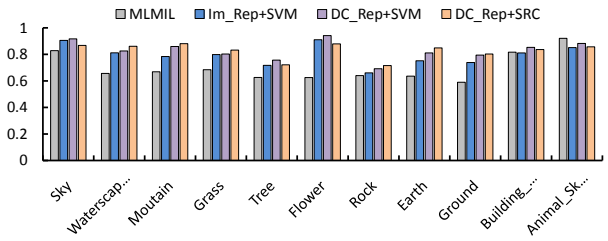

Figure 4. The AUCs of 11 classes of the Corel dataset using MLMIL, Im_Rep, Reg_Rep and our proposed DC_Rep.

ing samples when approximating the test image by linear combination, it could perform better with larger number of training images. The performances yielded by MLMIL on these datasets are lower than our proposed method. Furthermore, the JSEG segmentation used in MLMIL takes about 3 seconds for each image in the MSRC dataset, while our proposed method does not need segmentation for multi-label image annotation and can finish in about 0.01s (2.4GHz quad CUP and 3G RAM; not including feature extraction time).

Table 4. Label ranking WNDCGs on the MSRC dataset of different methods. Note that MLMIL yields 0.775.

Classifier	Im_Rep	Reg_Rep	DC_Rep
SVM	0.788	0.751	**0.832**
SRC	0.653	0.798	**0.820**

Table 5. Label ranking WNDCGs on the Corel dataset of different methods. Note that MLMIL yields 0.771.

Classifier	Im_Rep	DC_Rep
SVM	0.726	**0.766**
SRC	0.654	**0.767**

5.3. Label Ranking Results

Table 4 and Table 5 show the WNDCGs of label ranking using different methods on the two datasets. The proposed DC_Rep shows to be superior to others for its good representative and discriminative property. The similar background information does not help to rank amongst the labels that the image contains, therefore comparing to the multi-label classification, in this task the proposed DC_Rep achieves more improvement on the MSRC dataset. Again we can see Im_Rep does not coincide with the SRC, and for other representation methods SRC shows similar or better performance than SVM. Furthermore, SRC can reasonably predict the true number of relevant labels with its sparse solution while SVM can not. The proposed DC_Rep shows higher ranking performance than the other.

Figure 5 gives the multi-label annotation and ranking results for some sample images of various scenes from the MSRC dataset and the Corel dataset respectively. With the proposed DC_Rep and SRC, multi-label annotation and label ranking can be done simultaneously and the nonzero labels are automatically given. The results show to be promising and the proposed method can be useful for several applications.

6. Conclusions and Future Works

Conventional approaches to image decomposition heavily rely on the local appearance and neglect the context information across multiple images. We have developed a contextual image decomposition approach from the perspective of high-level vision tasks. The decomposed label representations can be naturally applied to multi-label image annotation and label ranking applications. It is worth noting that in contrast to many existing approaches to multi-label image annotation in which segmentation is a fundamental step, we can achieve multi-label annotation as well as label ranking without the employment of explicit image segmentation. The possible future works include the investigation of different visual representations in addition to the Bag-of-Word and applying the proposed approach to a larger-scale real-world dataset like Flickr [1].

(a) MSRC

(b) Corel

Figure 5. Sample results of multi-label image annotation and ranking on the two datasets using DC_Rep+SRC.

Acknowledgement

This work is partially supported by Singapore NRF/IDM Program, under research Grant NRF2008IDM-IDM004-029, partially supported by National Strategic R&D Program for Industrial Technology, and supported by the Agency for Defense Development for Unmanned Technology Research Center project.

References

[1] http://www.flickr.com/.

[2] M. Bhati. *Practical Optimization Methods with Mathematica Applications*. New York: Springer-Verlag, 2000.

[3] M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern Recognition*, 2004.

[4] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. *ICCV*, 2007.

[5] C. Chang and C. Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[6] Y. Deng and B. S. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Trans. on PAMI*, 23(8):800–810, 2001.

[7] X. He, R. Zemel, and M. A. Carreira-Perpinan. Multiscale conditional random fields for image labeling. *CVPR*, 2004.

[8] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 2002.

[9] T. Li and I. S. Kweon. A semantic region descriptor for local feature based image classification. *ICASSP*, 2008.

[10] X. Li, C. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. *ACM Multimedia Information Retrieval*, 2008.

[11] O. Maron, and A. L. Ratan. Multiple-Instance Learning for Natural Scene Classification. *ICML*, 1998.

[12] X. Ma and W. Grimson. Learning coupled conditional field for image decomposition with application on object categorization. *CVPR*, 2008.

[13] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. *ECCV*, 2006.

[14] J. Shi and J. Malik. Normalized cuts and image segmentation. *CVPR*, 1997.

[15] Z. Tu, X. Chen, A. Yuille, and S. Zhu. Image parsing: unifying segmentation, detection and recognition. *IJCV*, 2005.

[16] J. Wright, A. Yang, A. Ganesh, S. Shastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. on PAMI*, 2008.

[17] S. Yu, R. Gross, and J. Shi. Concurrent object recognition and segmentation by graph partitioning. *NIPS*, 2002.

[18] J. Yuan, J. Li, and B. Zhang. Exploiting spatial context constraints for automatic image region annotation. *ACM Multimedia*, 2007.

[19] Z. J. Zha, X. Hua, T. Mei, J. Wang, and Z. Wang. Joint multi-label multi-instance learning for image classification. *CVPR*, 2008.

[20] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 2007.

[21] Z. Zhou and M. Zhang. Multi-instance multi-label learning with application to scene classification. *NIPS*, 2007.