Retrieval-based Face Annotation by Weak Label Regularized Local Coordinate Coding *

Dayong Wang†, Steven C.H. Hoi†, Ying He†, Jianke Zhu‡
†School of Computer Engineering, Nanyang Technological University, Singapore.
‡College of Computer Science, Zhejiang University, Hangzhou, 310027, China.
{s090023, chhoi, yhe}@ntu.edu.sg, jkzhu@zju.edu.cn

ABSTRACT

Retrieval-based face annotation is a promising paradigm in mining massive web facial images for automated face annotation. Such an annotation paradigm usually encounters two key challenges. The first challenge is how to efficiently retrieve a short list of most similar facial images from facial image databases, and the second challenge is how to effectively perform annotation by exploiting these similar facial images and their weak labels which are often noisy and incomplete. In this paper, we mainly focus on tackling the second challenge of the retrieval-based face annotation paradigm. In particular, we propose an effective Weak Label Regularized Local Coordinate Coding (WLRLCC) technique, which exploits the local coordinate coding principle in learning sparse features, and meanwhile employs the graph-based weak label regularization principle to enhance the weak labels of the short list of similar facial images. We present an efficient optimization algorithm to solve the WLRLCC task, and develop an effective sparse reconstruction scheme to perform the final face name annotation. We conduct a set of extensive empirical studies on a large-scale facial image database with a total of $6,000$ persons and over $600,000$ web facial images, in which encouraging results show that the proposed WLRLCC algorithm significantly boosts the performance of the regular retrieval-based face annotation approaches.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Experimentation

Keywords

web facial images, auto face annotation, unsupervised learning

1. INTRODUCTION

In the digital era today, people can easily capture a photo by all kinds of digital devices, and share it through the internet by various online tools, such as web photo sharing portals (e.g., Flickr) or social networks (e.g., Facebook). Among vast digital images and photos shared on the internet, a considerable amount of them are related to human facial images because they are closely related to social activities of human beings. The rapid growth of facial images has created many research problems and opportunities for a variety of real-world applications. An important technique in this area is *automated face annotation*, which aims to tag human names to a novel unlabeled facial image automatically.

Auto face annotation is beneficial to a number of real-world applications. For example, using auto face annotation techniques, online photo sharing sites or social networks can automatically annotate users' uploaded photos to facilitate online photo search and management tasks. Auto face annotation can also be applied in news video domain to detect important persons appeared in the videos to facilitate news video retrieval and summarization tasks.

Typically, face annotation is formulated as an extended face recognition problem, in which face classification models are trained from a collection of well-controlled labeled facial images using supervised machine learning techniques [3, 25, 38, 41, 44]. Such approaches are referred to as "model-based face annotation" techniques. Despite being studied extensively, they suffer from several common drawbacks. First, it is usually time-consuming and expensive to collect a large amount of human-labeled training facial images. Second, it is difficult to generalize the models when new training data or novel persons are added, in which an intensive re-training process is usually required. Last but not least, the annotation or recognition performance often scales poorly when the number of persons/classes is very large.

Recent studies [33] have attempted to explore a promising retrieval-based annotation paradigm for facial image annotation by mining the world wide web (WWW), where a mass number of weakly labeled facial images are freely available. Instead of training explicit classification models by the regular model-based face annotation approaches, the retrieval-based face annotation (RBFA) paradigm aims to tackle the automated face annotation task by exploiting content-based image retrieval (CBIR) techniques [27, 37] in mining massive weakly labeled facial images on the web. Such a paradigm was somewhat inspired by the search-based image annotation techniques [34] for generic image annotation since face annotation can be generally viewed as a sub-problem of generic image annotation [10, 11, 29, 32, 36], which has been extensively studied but remains a very challenging open problem.

In general, there are two main challenges faced by the emerging retrieval-based face annotation (RBFA) technique. The first challenge is how to efficiently retrieve a short list of most similar facial images from a large facial image database, which typically relies on an effective content-based facial image retrieval solution. The

*Area chair: Lexing Xie

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

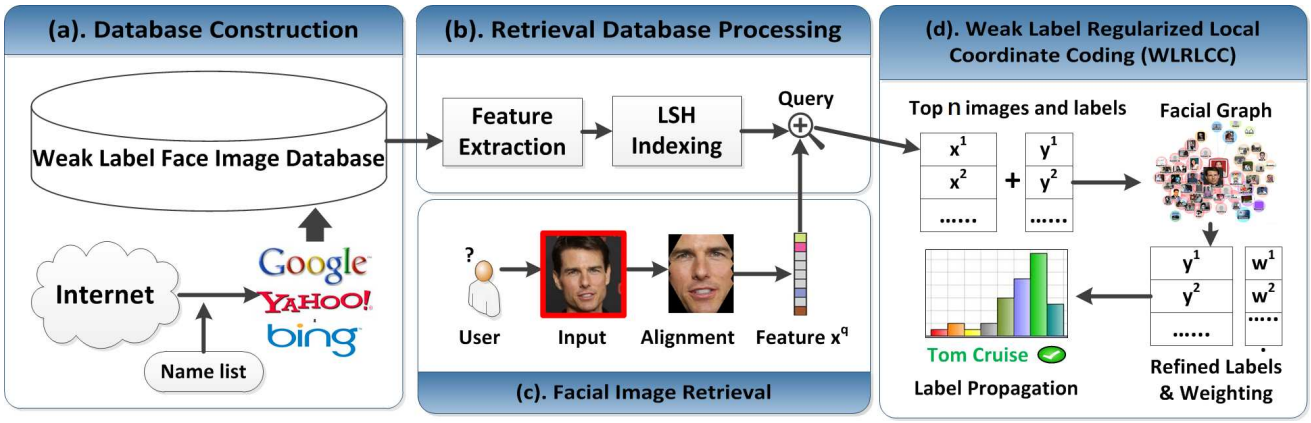


Figure 1: The system flow of the Retrieval-based Face Annotation (RBFA) scheme. (a) We collect weakly-labeled facial images from WWW using a web search engine; (b) We perform face detection and alignment, then extract GIST features from the detected faces, and finally apply LSH [9] to index the high-dimensional facial features; (c) A query facial image uploaded by the user is transformed into a feature vector with the same preprocessing step; using our content-based facial image retrieval engine, a short list of top-n most similar facial images and their associated names are retrieved and passed to the next learning and annotation stage; (d) The proposed WLRLCC scheme is applied to return the final list of (ranked) annotated face names.

second challenge is how to effectively exploit the short list of candidate facial images and their weak label information for the face name annotation task, which is critical because the associated labels of web facial images are often noisy and incomplete due to the nature of web images.

In this paper, we investigate the retrieval-based face annotation scheme by mainly addressing the second challenge. In particular, we propose a novel Weak Label Regularized Local Coordinate Coding (WLRLCC) technique to tackle the retrieval-based face annotation problem, which attempts to boost the annotation performance by a unified learning scheme, which exploits the local coordinate coding in achieving more effective features and makes use of the graph-based regularization to enhance the weak label simultaneously. In particular, given a query facial image for annotation, we firstly search for a short list of top-n most similar facial images from a weakly labeled web facial image database. After that, we apply the proposed WLRLCC algorithm to obtain a more discriminative local coordinate coding representation and an enhanced label matrix as well. Finally, we apply an effective sparse reconstruction scheme for the final facial name annotation. As a summary, the main contributions of this paper include:

- We propose a novel Weak Label Regularized Local Coordinate Coding (WLRLCC) technique, which constructs an effective sparse coding and enhances the weak label by exploiting both the local coordinate coding and graph-based weak label regularization principles.
- We propose an efficient algorithm to solve the proposed WLRLCC problem, and conduct extensive experiments to evaluate the proposed algorithm for automated face annotation on a large-scale web facial image database.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 briefly introduces the proposed retrieval-based face annotation (RBFA) framework. Section 4 presents the proposed Weak Label Regularized Local Coordinate Coding (WLRLCC) scheme together with an effective face name annotation solution based on sparse reconstruction. Section 5 shows the experimental results of our empirical studies. Finally, Section 6 discusses the limitations, and Section 7 concludes this paper.

2. RELATED WORK

Our work is closely related to several groups of research work.

The first group is the studies of face detection, verification, and recognition. Comprehensive surveys on such topics can be found in [42, 14]. Although traditional face recognition methods can be extended for auto face annotation, they usually suffer from some common drawbacks. For example, they often require high-quality facial image databases collected in well-controlled environments. This drawback has been partially improved by some recent work on benchmark studies of unconstrained face detection and verification techniques on facial images collected from the web, such as the LFW benchmark [16, 21, 4]. However, traditional face recognition techniques have to train explicit classification models, which usually scale poorly when the number of persons/classess is large and are nontrivial to re-train the models when adding new training data or additional persons.

The second group is the studies of face annotation on collections of personal/family photos. Several studies [30, 7, 8] have mainly focused on the annotation task on collections of personal/family photos, which often contain rich context clues, such as personal/family names, social context, GPS tags, timestamps, etc. In addition, the number of persons/classess is usually quite small, making such annotation tasks less challenging. These techniques usually achieve fairly impressive annotation results, and some techniques have been successfully deployed in commercial applications, e.g., Apple iPhoto and Microsoft easyAlbum [8].

The third group is the studies of face annotation in mining weakly labeled facial images on the web. Some studies consider a human name as the input query, and mainly aim to refine the text-based search results by exploiting visual consistency of facial images, which is closely related to automatic image re-ranking problems. For example, [22] proposed a graph-based model for finding the densest sub-graph as the most related result. Le and Satoh [17] proposed a new local density score to represent the importance of each returned images. Unlike these studies of filtering the text-based retrieval results, some studies have attempted to directly annotate each facial image with the names extracted from its caption information. For example, [2] proposed a possibility model combined with a clustering algorithm to estimate the relationship between the

facial images and the names in their captions. Finally, our work is closely related to some recent studies on the retrieval-based face annotation paradigm [33], which have attempted to attack the automated face annotation problem by exploiting content-based facial image retrieval techniques in mining web facial images. Some recent work [37] mainly addressed the first challenge of the RBFA paradigm, in which an effective image representation has been proposed by employing both the local and global features. Another recent work is the unsupervised label refinement (ULR) technique in [33] to address the second challenging problem where the initial weak label is enhanced by mining the facial graph information over the whole retrieval database. The work in the paper differs from the ULR scheme in several aspects. First, the ULR scheme is difficult to handle huge databases due to its high computational cost, while the proposed WLRLCC algorithm is applied only to a short list of most similar images. Second, the simple majority voting scheme of the ULR scheme might not be effective enough in exploiting the short list of most similar images; in contrast, the proposed WLRLCC algorithm comprehensively resolves this problem by fully exploiting the short list of top similar images via a unified optimization scheme of learning both sparse features and enhanced labels.

The fourth group is the studies of generic image annotation. The commonly studied techniques usually apply existing object recognition techniques to train classification models from human-labeled training images or attempt to infer the correlations probabilities between query images and annotation keywords [10, 11, 5, 13]. In recent years, the search-based image annotation paradigm in mining web images has attracted more and more research attention [34, 24, 31, 23]. Some studies in this area have attempted to develop efficient content-based indexing and searching techniques to facilitate the annotation/recognition tasks. For example, Russell et al. [24] developed a large collection of web images with ground truth labels to facilitate object recognition research. Wang et al. [34] proposed an efficient search-based annotation scheme for auto image annotation by combining both textural and visual information. There are also some studies that aim to address the final annotation process with effective label propagation. For example, Wright et al. [35] proposed a classification algorithm based sparse representation, which predicts the label information based on the class-based feature reconstruction. Tang et al. [29] presented a sparse graph-based semi-supervised learning (SGSSL) approach to annotate web images. Wang et al. [32] proposed another sparse coding based annotation framework (MSR), where the label-based graph is used to learn a linear transformation matrix for feature dimension reduction, and sparse reconstruction is employed for the subsequent label propagation step. Wu et al. [36] proposed to select heterogeneous features with structural Grouping Sparsity and suggested a Multi-label Boosting scheme (MtBGS) for feature regression, in which a group sparse coefficient vector is achieved for each class (category) and further used for predicting new instances. Wu et al. [37] proposed a multi-reference re-ranking (MRR) scheme for improving the retrieval process.

Finally, we note that the proposed learning methodology for WLRLCC is partially inspired by some existing works in machine learning, including local coordinate coding [40, 15], graph-based semi-supervised learning [45, 6] and multi-label learning [28].

3. RETRIEVAL-BASED FACE ANNOTATION FRAMEWORK

In this section, we briefly introduce the proposed Retrieval-based Face Annotation (RBFA) paradigm. Figure 1 illustrates the proposed framework, which consists of the following four major stages: (i) data collection of facial images from WWW, (ii) facial image preprocessing and high-dimensional facial feature indexing, (iii) content-based facial image retrieval for a query facial image, (iv) face annotation by the proposed Weak Label Regularized Local Coordinate Coding (WLRLCC) algorithm. The details of each stage are described as follows.

The first stage, as shown in Figure 1(a), is to collect a database of weakly labeled facial images, which can be crawled from WWW. In particular, we can choose a list of desired human names and submit them to existing web search engines (e.g., Google) for crawling their related web facial images. As the output of this crawling process, we obtain a collection of web facial images, each of them is associated with some human names. Given the nature of web images, these facial images are usually noisy, which may be incorrect or incomplete. We thus refer to such web facial images with noisy names as weakly labeled facial images.

The second stage, as shown in Figure 1(b), is to pre-process the weakly label facial image database, including face detection, face alignment, facial feature representation, and high dimension feature indexing. For facial region detection and alignment, we adopt OpenCV and the unsupervised face alignment technique proposed in [43]. For facial feature representation, we extract the GIST features [26] to represent the extracted facial regions. Finally, we apply the Locality-Sensitive Hashing (LSH) [9] to index the GIST features in our solution.

The previous two stages must be done before annotating a query facial image. The next two stages are related to online processes of annotating a query facial image. As shown in Figure 1(c), given a query facial image, we conduct a similar face retrieval process to find a short list of most similar faces (e.g., top-n similar faces) from the indexed face databases using the LSH technique.

After obtaining the top-n most similar faces, the last stage is to apply the proposed Weak Label Regularized Local Coordinate Coding (WLRLCC) algorithm for name annotation, as shown in Figure 1(d). Specifically, the proposed WLRLCC algorithm learns local coordinate coding for each of the similar facial images and enhances the weak label matrix via an iterative optimization process. Based on the learning results, a sparse reconstruction algorithm is applied to perform the final face name annotation. Next we present the details of the proposed WLRLCC technique.

4. WEAK LABEL REGULARIZED LOCAL COORDINATE CODING (WLRLCC)

In this section, we present the proposed Weak Label Regularized Local Coordinate Coding algorithm for the face annotation task based on a list of top-n most similar facial images. We first introduce some preliminaries and notations, followed by the problem formulation and the proposed algorithms, respectively. Finally, we discuss the sparse reconstruction method for the final name annotation.

4.1 Preliminaries

Consider a query facial image $x^q \in \mathbb{R}^d$, a d-dimension feature vector with some unknown class label denoted as y^q. The goal of a retrieval-based face annotation task is to estimate y^q based on the retrieval result from a large facial image database. Assume the short list of top-n most similar facial images to x^q are $\{(x^i, y^i)_{i=1}^n\}$, where $y^i \in \{1, 2, \ldots, m\}$ is the name label of its corresponding facial image x^i, and m is the total number of classes (names) among all the top-n facial images.

Let us denote by the feature matrix of the retrieval results $X =$

$[x^1, x^2, \ldots, x^n]$. We represent the initial name information with a class-membership-indicator matrix $\tilde{Y} \in \mathbb{R}^{n\times m}$, where $\tilde{Y}_{i\star}$, the i-th row of matrix, denotes the class label vector for x^i, $\tilde{Y}_{ij} = 1$ if $j = y^i$, and $\tilde{Y}_{ij} = 0$ otherwise. Due to the nature of web images, the initial "weak" label information $\tilde{Y}$ is often noisy and incomplete. Thus, it is critical to find some effective solution to enhance the initial weak label information. Finally, it is common to know that facial images usually follow some manifold structure, which motivates us to apply the local coordinate coding [40] to find a more effective sparse representation.

4.2 Problem Formulation

4.2.1 Sparse Features via Local Coordinate Coding

Sparse representation has been successfully applied in many applications, e.g., face recognition [35], graph construction [29], and feature selection [36], etc. Let us denote by $x \in \mathbb{R}^d$ some observed feature vector for reconstruction, and consider an under-determined system of linear equation as follows:

$$x = [D, I] \begin{bmatrix} s \\ \xi \end{bmatrix} = D \cdot s + I \cdot \xi,$$

where D is the overcomplete dictionary, I is an identity matrix, and ξ is a noise term. If the solution s is sparse enough, it should be recovered by solving the following convex optimization problem:

$$\min_{[s;\xi]} \sigma_{\mathtt{SC}}([s;\xi]), \quad \text{s.t.} \quad x = D \cdot s + I \cdot \xi \tag{1}$$

where the sparsity penalty function is defined as $\sigma_{\mathtt{SC}}(\cdot) = \|\cdot\|_1$. We can apply the same idea to reconstruct the sparse representation s^i of the i-th facial image x^i in the top-n retrieval results based on the dictionary $B = [X, I] \in \mathbb{R}^{d\times(n+d)}$ as follows:

$$\hat{s}^i = \arg\min_{\hat{s}} \frac{1}{2}\|x^i - B\hat{s}\|^2 + \lambda \cdot \sigma_{\mathtt{SC}}(\hat{s}), \text{ s.t. } \hat{s}_i = 0 \tag{2}$$

where $\hat{s}^i = [s^i; \xi^i]$, s^i is the sparse coding of x^i, and ξ^i is the coefficient related to the noise of the facial image. The constraint $\hat{s}_i = 0$ is to avoid using x^i itself for reconstruction.

Following the Local Coordinate Coding (LCC) [40], it is important to exploit the locality information for the linear embedding of high dimensional data on the manifold. In contrast to traditional sparse coding, the intuition of LCC is to assign a larger coefficient to the dictionary item that is closer to the encoding item. Specifically, instead of adopting the sparsity term $\sigma_{\mathtt{SC}}(\hat{s})$ directly, a locality weight is assigned to each element in $\hat{s}$. Thus, the regularizer of LCC is formulated as follows:

$$\sigma_{\mathtt{LCC}}(\hat{s}) = \sum_{k=1}^{n+d} |\hat{s}_k| \cdot \|B_{\star k} - x^i\|^2 \tag{3}$$

where $B_{\star k}$ is the k-th column of dictionary B, and $\hat{s}_k$ is the k-th element of $\hat{s}$.

The regularizer $\sigma_{\mathtt{LCC}}(\hat{s})$ in LCC has a close relationship with the non-negative sparse coding (NNSC) [15], in which all the elements in $\hat{s}$ are forced to be non-negative:

$$\sigma_{\mathtt{NNSC}}(\hat{s}) = \sum_k \hat{s}_k, \text{ s.t. } \hat{s} \geq 0 \tag{4}$$

In the above regularizer, because of the non-negative constraint, a dictionary item that is close to the input item x^i (i.e., small $\|B_{\star k} - x^i\|^2$) is more likely to be active (nonzero) in NNSC, which in some sense is similar to LCC. By this motivation, we propose to combine the two regularizers to strengthen the locality constraint, which will also simplify our optimization problem. As a result, we can reformulate the optimization of learning the sparse representation for x^i as follows:

$$\begin{gathered} e(\hat{s}; x^i) = \min_{\hat{s}} \frac{1}{2}\|x^i - B\hat{s}\|^2 + \lambda \sum_{k=1}^{n+d} \hat{s}_k \cdot \|B_{\star k} - x^i\|^2 \\ \text{s.t. } \hat{s}_k \geq 0, k = 1, 2, \ldots, n+d \text{ and } \hat{s}_i = 0. \end{gathered} \tag{5}$$

Finally, we can give the formulation of the non-negative local coordinate coding problem for all the facial images in the top-n retrieval results:

$$E_1(\hat{S}; X) = \sum_{i=1}^{n} e(\hat{s}; x^i). \tag{6}$$

where $\hat{S} \in \mathbb{R}^{(n+d)\times n} = [S; \Xi]$, $S \in \mathbb{R}^{n\times n}$ is the local coordinate coding of X, and $\Xi \in \mathbb{R}^{d\times n}$ is the noise matrix.

4.2.2 Weak Label Enhancement

The previous formulation shows that the j-th local sparse coefficient s^i_j of facial image x^i essentially encodes the locality information between x^i and $x^j, j \neq i$. Specifically, a larger value of s^i_j indicates that x^j is more representative of x^i, i.e., having a larger contribution to the reconstruction of x^i. In addition, from a view of graph-based semi-supervised learning, for any two facial images, the smaller their local distance, the more likely they should belong to the same person. As a result, a larger value of s^i_j implies that the name labels of x^i and x^j are more likely to be the same.

Based on the above motivation, we can give the following formulation to enhance the initial weak label matrix $\tilde{Y}$ as follows:

$$E_2(Y, S) = \min_{Y\geq 0} \frac{1}{2} \sum_{i,j} s^i_j \|Y_{i\star} - Y_{j\star}\|^2 + \lambda \|(Y - \tilde{Y}) \circ M\|^2_{\mathtt{F}} \tag{7}$$

M is a "sign" matrix $M = [sign(\tilde{Y}_{ij})]$ where $sign(x) = 1$ if $x > 0$ and 0 otherwise, and $\circ$ denotes the Hadamard product of two matrices. In the above objective function, the first term enforces that the class labels of two facial images x^i and x^j to be similar if the local sparse coefficient s^i_j is large, and the second term is a regularization term that prevents the refined label matrix being deviated too much from the initial weak matrix. Since the initial label matrix is noisy and incomplete, we apply the regularization of the second term on only these nonzero elements in $\tilde{Y}$.

In general, the optimal solution to the problem in Eq. (7) is dense; however, the ideal true label matrix is often very sparse. Following the suggestion in [33], we introduce some convex sparsity constraints to take into the consideration of sparsity, i.e., $\|Y_{i\star}\|_1 \leq \varepsilon, \varepsilon \geq 1$, where $i = 1, 2, \ldots, n$ (we choose $\varepsilon = 1$ in this work). These constraints are included to limit the number of name labels assigned to each facial image.

4.2.3 Weak Label Regularized Local Coordinate Coding

The above two optimization tasks of "sparse feature learning" and "label enhancement" are performed separately. In particular, the sparse features S are first learned from the optimization in Eq. (5), and then used by the optimization in Eq. (7) to refine the label matrix Y. To better exploit the potential of the two learning approaches, we propose the Weak Label Regularized Local Coordinate Coding (WLRLCC) scheme, which aims to reinforce the two learning tasks via a unified optimization framework. Specifically, the optimization of WLRLCC can be formulated as follows:

$$Q(\hat{S},Y)=E_1(\hat{S};X)+E_2(Y,S)=\min_{\hat{S},Y}\frac{1}{2}\|B\hat{S}-X\|_{\mathrm{F}}^2+$$
$$\lambda_1 tr(\mathbf{1}\cdot(\hat{S}\circ V))+\lambda_2 tr(Y^\top LY)+\lambda_3\|(Y-\tilde{Y})\circ M\|_{\mathrm{F}}^2 \quad (8)$$
$$\text{s.t. } \hat{S}_{ii}=0, \|Y_{i\star}\|_1\leq 1, i=1,2,\ldots,n,\ \hat{S}\geq 0, Y\geq 0$$

where $V\in\mathbb{R}^{(n+d)\times n}$, $V_{ij}=\|B_{\star i}-X_{\star j}\|^2$, $L=D-S$, D is a diagonal matrix, with $D_{ii}=\frac{\sum S_{i\star}+\sum S_{\star i}}{2}$, $Y\in\mathbb{R}^{n\times m}$, $\mathbf{1}$ is all-one-element matrix with dimension $n\times(n+d)$, $\circ$ denotes the Hadamard product of two matrices, and $tr(\cdot)$ denotes a trace function. In the above, $\lambda_2 tr(Y^\top LY)$ is a label smoothness regularizer which connects between the label matrix and the sparse features.

Remark. Although WLRLCC is similar to some existing sparse representation schemes, it differs from the existing approaches in two major aspects. First, instead of using the traditional sparse representation for semantic graph construction, a non-negative local coordinate coding algorithm is adopted, which exploits the advantages of both NNSC [15] and LCC [40]. Second, WLRLCC employs a unified optimization scheme to solve both sparse feature learning and label enhancement in an iterative approach; as a result, the enhanced label information is also exploited by the sparse coding step, which in some sense becomes a weakly supervised sparse learning approach while traditional sparse coding is often unsupervised.

4.3 Optimization

The optimization problem in Eq. (8) is generally non-convex. To solve this challenging optimization, we propose to solve $\hat{S}$ and Y alternatively by iteratively solving two optimization steps: (1) *Code Learning*, and (2) *Label Learning*. The step of updating $\hat{S}$ can be transformed into a weighted non-negative sparse coding problem, and the step of updating Y is a quadratic programming problem.

4.3.1 Code Learning

By first fixing the label matrix Y and ignoring the constant terms, the optimization problem in Eq. (8) can be reformulated as follows:

$$Q_Y(\hat{S})=\min_{\hat{S}}\frac{1}{2}\|B\hat{S}-X\|_{\mathrm{F}}^2+tr(\mathbf{1}\cdot(\hat{S}\circ Z))$$
$$\text{s.t. } \hat{S}_{ii}=0, i=1,2,\ldots,n, \text{and } \hat{S}\geq 0 \quad (9)$$

where $Z=\lambda_1 V+\lambda_2 U$, $U\in\mathbb{R}^{(n+d)\times n}$, for all $j=1,2,\ldots,n$, if $i\leq n$, $U_{ij}=\frac{1}{2}\|Y_{i\star}-Y_{j\star}\|^2$; otherwise $U_{ij}=0$. The other variables follow the same definitions in Eq.(8).

The optimization problem in Eq. (9) can be further separated into a series of sub-problems for each coding coefficient $\hat{S}_{\star i}$ of facial image $X_{\star i}$. Each sub-problem is a weighted non-negative sparse coding problem, which can be written into the following optimization:

$$\min_{\hat{s}\geq 0}\frac{1}{2}\|B\hat{s}-X_{\star i}\|^2+\lambda_1\sum_{k=1}^{n+d}\hat{s}_k\cdot Z_{ki} \quad \text{s.t. } \hat{s}_i=0. \quad (10)$$

In our approach, we adopt the Fast Iterative Shrinkage and Thresholding Algorithm (FISTA) [1], a popular and efficient algorithm for the linear inverse problem that has been already implemented for sparse learning in [19]. Since we aim to solve the problem related to only the top-n images in the retrieval results, where n is usually small, FISTA is efficient enough for our application.

4.3.2 Label Learning

Similarly, by fixing $\hat{S}$ and ignoring the constant terms, the optimization problem in Eq.(8) can be reformulated as follows:

$$Q_{\hat{S}}(Y)=\min_{Y} tr(Y^\top LY)+\lambda\|(Y-\hat{Y})\circ M\|_{\mathrm{F}}^2$$
$$\text{s.t. } Y\geq 0, \|Y_{i\star}\|_1\leq 1, i=1,2,\ldots,n. \quad (11)$$

where $\lambda=\frac{\lambda_3}{\lambda_2}$, and the other variables follow the same definitions in Eq.(8). The optimization problem is a quadratic program with a series of closed and convex constraints (ℓ_1 ball). In order to efficiently solve this problem, we first vectorize the label matrix $Y\in\mathbb{R}^{n\times m}$ into a column vector $\bar{y}=vec(Y)\in\mathbb{R}^{(n\cdot m)\times 1}$, and then rewrite the previous optimization problem as follows:

$$Q(\bar{y})=\min_{\bar{y}\geq 0}\bar{y}^\top\Psi\bar{y}+\mathbf{c}^\top\bar{y},\ \text{s.t. } \forall i,\sum_{k=0}^{m-1}\bar{y}_{k\cdot n+i}\leq 1. \quad (12)$$

where $\Psi=I_m\otimes L^\top+\lambda R$, I_m is an identity matrix with dimension $m\times m$, $R=diag(vec(M))$, $\mathbf{c}=-2\lambda R^\top\cdot vec(\tilde{Y})$, and $\bar{y}_j$ is the j-th element in $\bar{y}$. In order to solve this problem, we employ the multi-step gradient scheme [1] and the efficient Euclidean projection algorithm [20].

Specifically, in the multi-step gradient scheme, in order to achieve the optimal solution $\bar{y}^\star$, we recursively update two sequences $\{\bar{y}^{(k)}\}$ and $\{z^{(k)}\}$, where k is the iteration step. Commonly at each iteration k, the variance $z^{(k)}$ is named as the search point and used to construct the combination of the two previous approximate solutions $\bar{y}^{(k-1)}$ and $\bar{y}^{(k-2)}$. In our problem, the sub-block problem is defined as:

$$\bar{y}^{(k+1)}=\arg\min_{\bar{y}\geq 0}\frac{t}{2}\|\bar{y}-\mathbf{v}\|^2,\ \text{s.t. } \forall i,\sum_{k=0}^{m-1}\bar{y}_{k\cdot n+i}\leq 1. \quad (13)$$

where $\mathbf{v}=z^{(k)}-\frac{1}{t}\mathbf{g}$ and $\mathbf{g}=2\Psi z^{(k)}+\mathbf{c}$, t could be a fixed positive constant or searched in a backtracking step [1]. The key problem is to solve Eq. (13) efficiently. In our work, we employ the Euclidian projection algorithm [20], which has been shown as an efficient solution to such kind of ℓ_1 ball constrained problem with linear time complexity of $O(n)$ as compared with other algorithms of $O(n\log(n))$.

Further, in order to apply the Euclidian projection algorithm, we split $\bar{y}$ into a series of m-dimension sub-vectors with $\bar{y}^i=[\bar{y}_{k\cdot n+i})]_{k=0}^{m-1}$, where $i=1,2,\ldots,n$, and similarly we can split the vector $\mathbf{v}$. Specifically, each optimal solution $\bar{y}^{i\star}$ for sub-vector $\bar{y}^i$ can be achieved by solving the following problem:

$$\bar{y}^{i\star}=\arg\min_{\bar{y}^i\geq 0}\frac{t}{2}\|\bar{y}^i-v^i\|^2,\quad \text{s.t. } \|\bar{y}^i\|_1\leq 1. \quad (14)$$

The problem in Eq. (14) has a non-negative constraint, which is different from the one in [20]. The optimal solution to this problem is given as:

$$\bar{y}_j^{i\star}=sign(v_j^i)\cdot\max(|v_j^i|-\lambda^\star,0), j=1,2,\ldots,m. \quad (15)$$

where $sign(\cdot)$ is the sign function in Eq. (7), $\bar{y}_j^i$ and v_j^i are the j-th elements in $\bar{y}^i$ and v^i, and $\lambda^\star$ is the optimal solution for the dual form of Eq. (14). Suppose $S=\{j|v_j^i\geq 0\}$, the value of $\lambda^\star$ can be computed as follows:

$$\lambda^\star=\begin{cases}0 & \sum_{k\in S}v_k^i\leq 1,\\ \bar{\lambda} & \sum_{k\in S}v_k^i>1.\end{cases} \quad (16)$$

where $\bar{\lambda}$ is the unique root of function $f(\lambda)=\sum_{k\in S}\max(|v_k^i|-\lambda,0)-1$, which is continuous and monotonically decreasing in $(-\infty,\infty)$. The root $\bar{\lambda}$ can be achieved by a bisection search in

linear time complexity of $O(dim(\bar{y}^i))$. It is well-known that the multi-step gradient scheme converges at $O(\frac{1}{k^2})$, where k is the iteration step, indicating our optimization can be solved efficiently.

4.4 Face Name Annotation

After obtaining the sparse coding matrix S and the enhanced label matrix Y, the last key step of our framework is to perform the face name annotation. First, the query facial image x^q is also projected into the local coordinate coding space. In particular, the new coding s^q w.r.t. x^q can be achieved by solving the following optimization problem:

$$[s^q;\xi^q] = \arg\min_{\hat{s}\geq 0} \frac{1}{2}\|B\hat{s} - x^q\|^2 + \lambda \sum_k \hat{s}_k \cdot \|B_{\star k} - x^q\|^2 \quad (17)$$

where the parameter setting is the same as the previous WLRLCC algorithm.

Although the manifold structure is supposed to be well fitted in the new local coordinate space, it is still unsuitable to directly use a one-vs-one similarity measure for face name annotation. For the second step, we employ a sparse reconstruction scheme in the local coordinate coding space to recover the potential weighting vector w^q for name annotation in the label space. The optimization problem is given as follows:

$$\min \|w^q\|_1 \quad \text{s.t. } s^q = S_{I_n} \cdot w^q \quad (18)$$

where $S_{I_n} = S + I_n$, S is the local coordinate codes obtained from the previous step and I_n is an $n \times n$ identity matrix.

Finally, the label vector y^q can be directly computed by:

$$y^q = Y^\top \cdot w^q$$

The value $y^q_k, k \in \{1, 2, \ldots, m\}$ measures the confidence of the k-th name being assigned to the query facial image x^q; as a result, a name with a large confidence value will be ranked on the top position for the final annotation result.

5. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed Weak Label Regularized Local Coordinate Coding (WLRLCC) algorithm, we conduct an extensive set of experiments to compare WLRLCC with other annotation algorithms on a large real-world weakly labeled facial images database. In the following subsections, we first briefly introduce the construction of our weak labeled facial image database. We then discuss the compared algorithms and their parameter settings. Finally, we present and discuss the experimental results.

5.1 Experimental Testbed

To build our experimental testbed, we first collected a name list consisting of 6000 popular actor and actress names downloaded from the **IMDb** website http://www.imdb.com. We collected those names with the billboard: "Most Popular People Born In **yyyy**" of **IMDb**, where **yyyy** is the born year. e.g., the webpage [1] presents all the actors and actresses who were born in 1975 in the popularity order. Our name list covers the actors and actresses who were born between 1950 and 1990. We submitted each name from the list as a query to search for related web images by Google image search engine. The top 200 retrieved web images were crawled automatically. After that we adopt the OpenCV toolbox to detect the faces and adopt the Deformable Lucas-Kanade (DLK) algorithm [43] to align facial images into the same well-defined position. The non-face-detected web images were ignored directly. As a result, we collected over $600,000$ facial images in our database. We refer to this database as the "retrieval database", which will be used for facial image retrieval during the auto face annotation process.

In order to evaluate the effect of different database sizes, we also built three subsets of the whole database by including the images belonging to a subset of P most popular persons. For example, when choosing $P = 400$, we will choose the top $\frac{P}{40} = 10$ names from the name list of each year. We denote such a new database as "GDB-040K", which means that there are about over $40,000$ images in this sub-database. Following the same method, we totally construct four databases in different sizes, i.e., GDB-040K, GDB-100K, GDB-200K, GDB-600K.

We also built a "test dataset" by randomly choosing 120 names from our name list. We submitted each of these names as a text query to Google image search and crawled about 200 images from the top 200-th to 400-th search results. Note that we did not consider the top 200 retrieved images since they had already been collected in the retrieval database. This aims to examine the generalization performance of the proposed technique for unseen facial images. Since these facial images are often noisy, in order to form the ground truth labels for the test dataset, we requested our staff to manually examine the retrieved facial images and remove those irrelevant facial images for each name. As a result, the test database consists of about $1,600$ facial images with on average about 10 to 15 facial images for each person.

5.2 Compared Algorithms and Metric

To evaluate the performance of the proposed WLRLCC algorithm, we compare it against a number of different algorithms, including a common baseline based on weighted majority voting and five state-of-the-art algorithms for face/image annotation. Most of these algorithms were proposed in recent years, which have been briefly introduced in Section 2. As a fair comparison, we adopted the same GIST features [26] to represent all the facial images. To evaluate the annotation performances, we adopted the *hit rate* at the top-t annotated results as the performance metric, which measures the likelihood of having the true label among the top-t annotated names for a query facial image. The list of compared algorithms and their parameter settings are discussed as follows:

- "SMW": a baseline algorithm that simply adopts the soft-max weighted majority voting, which assigns a weighting coefficient w_i to the label vector $\tilde{Y}_{i\star}$ by $w_i = \frac{s(x^q,x^i)}{\sum_{j=1}^n s(x^q,x^j)}$, where $s(\cdot,\cdot)$ is defined as $s(x^q,x^i) = \frac{1}{1+\exp(-\|x^q-x^i\|^2)}$.
- "S-Recon": the sparse label reconstruction algorithm [35]. The label of x^q is predicted as $y^q = Y^\top \cdot s^q$, where s^q is the sparse representation of x^q achieved with Eq.(1).
- "SGSSL": the kNN Sparse Graph-based Semi-Supervised Learning algorithm [29], where parameters are set according the ones given by the authors.
- "MRR": the Multi-Reference Re-ranking algorithm [37]. The parameters setting follows the authors' suggestions of $N_p = 320$ and $\alpha = 8$. After the re-ranking step, the SMW algorithm is employed for the final name annotation.
- "MtBGS": the Multi-label Boosting by selecting heterogeneous features with structural Grouping Sparsity [36]. The parameters λ_1 and λ_2 are tuned via cross validation. As only one kind of feature is used, the number of groups in MtBGS is set to 1 in our experiment.
- "MSC" : the multi-label sparse coding framework, in which label information is used to build the semantic graph for feature embedding [32]. The parameter λ in the graph construc-

[1] http://www.imdb.com/search/name?birth_year=1975

tion and label propagation steps are determined via cross validation. The parameter β is set to 0.1 according to the paper.

- "WLRLCC": the proposed Weak Label Regularized Local Coordinate Coding algorithm. The parameter λ_2 in Eq. (8) is related to the *Code learning* step and *Label learning* step, where its contribution also depends on the setting of λ_1 and λ_3, respectively. In our experiments, we set λ_2 to 0.1, and tune λ_1 and λ_3 via cross validation. For the parameter in Eq. (17), we set it to the same value as λ_1.

Furthermore, we discuss the cross validation issue. For all the compared algorithms, we only conducted it in the experiments on the GDB-040K; as a result, the parameters with the best average hit rate results are directly used in the following experiments based on GDB-100K, GDB-200K, and GDB-600K, which is also helpful to evaluate the parameter sensitivity. In particular, for the experiments on GDB-040K, we randomly divide the test dataset into two parts of equal size, in which one part is used as the validation set to find the optimal parameters by a grid search, and the other part is used for performance evaluation. Such a procedure is repeated 10 times and the average performance is computed over the 10 trials, in which the parameters with the best average hit rate performance are recorded. For the experiments on GDB-100K, GDB-200K and GDB-600K, we also split the test database into two parts, and randomly selected one for performance evaluation with the previous recorded parameters. For all the algorithms, we report both the average performances and their standard deviations.

5.3 Impact of Top-n and Top-t Settings

Top-n	t=1	t=3	t=5	t=7	t=9	t=10
10	**0.570**	0.731	0.773	0.793	0.804	0.808
20	0.529	0.703	0.779	0.798	0.810	0.817
30	0.511	0.683	0.760	0.800	0.818	0.822
40	0.497	0.669	0.740	0.784	0.822	**0.832**
50	0.470	0.659	0.737	0.769	0.804	0.818
100	0.413	0.605	0.693	0.748	0.777	0.792
200	0.362	0.552	0.638	0.692	0.738	0.756

Table 1: The hit rate performance of baseline algorithm SMW with different settings of Top-n faces and Top-t names.

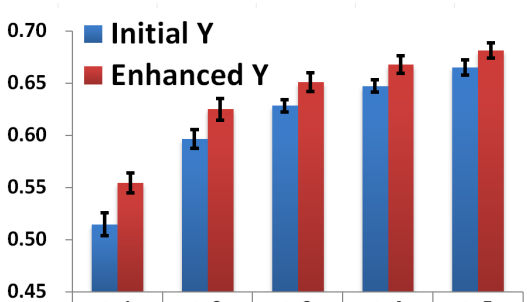
This experiment aims to evaluate the impact on the final annotation performance by varied settings of n and t values for Top-n retrieved facial images and Top-t annotated names. Table 1 shows the annotation performance of the baseline SMW algorithm with different settings of n and t values on the GDB-040K database. In this experiment, the initial weak label $\tilde{Y}$ is used as we aim to examine the relationship between n and t in the original database.

Several observations can be drawn from the experimental results. First of all, when n is fixed, increasing the value of t generally leads to better annotation performance. However, the improvement becomes marginal when t is larger than some threshold. Second, when increasing the value of n, the annotation performance decreases for a small t value (e.g., Top-t=1), but increases for a large t value (e.g., Top-$t = 10$). However, when n is large enough (i.e., Top-$n > 40$), increasing n definitely leads to degrade the annotation performance for all the t values. Finally, we found that a small n value often results in a high precision and a low recall of any top-t retrieval results, while a large n value often produces a high recall and a low precision.

The previous observations are beneficial to determining appropriate parameters in our experiments. Since SMW is a simple weighted majority voting algorithm without any re-ranking, it prefers the retrieval result with a high precision; thus its parameter n for Top-n retrieved images is set to 10. On the other hand, the other algorithms usually carrying a re-ranking or sample selection step typically prefer a high recall of the retrieved results such that a potentially larger pool of facial images can be exploited for re-ranking. Thus, the parameter n for all the other algorithms is set to 40.

5.4 Evaluation of Label Enhancement

This experiment aims to evaluate the performance of the refined label matrix Y. To ease our discussion, we only present the results of the proposed WLRLCC algorithm on the most noisy GDB-600K database. Similar observations can also been observed on the other databases by employing different algorithms. The WLRLCC algorithm is firstly applied to learn the local coordinate coding and enhance the label matrix. After that, both the refined label matrix Y and the initial weak label matrix $\tilde{Y}$ are used in the face name annotation step. The comparison results are presented in Figure 2.



	t=1	t=2	t=3	t=4	t=5
Initial Y	0.5145	0.5961	0.6283	0.6471	0.6649
Enhanced Y	0.5544	0.6248	0.6509	0.6679	0.6813

Figure 2: Evaluation of the enhanced label matrix Y on the GDB-600K database.

From the experimental results, it is obvious to observe that the enhanced label matrix Y significantly boosts the annotation performance, especially for small t values. This means that the proposed WLRLCC algorithm can efficiently exploit the graph-based locality information to refine the initial weak label matrix. In the following experiments, in order to evaluate the encoding results of the proposed WLRLCC algorithm, we will directly adopt the refined label matrix Y for all the compared algorithms to enable a fair comparison.

5.5 Evaluation of Auto Face Annotation

In this experiment, we compare the proposed WLRLCC algorithm with other algorithms on the GDB-040K database. Figure 3 and Table 2 show the average annotation performance at different t values, in which both mean and the standard deviation are reported. We can draw some observations for the results.

First of all, it is clear that the proposed WLRLCC algorithm consistently performs better than the other algorithms, especially for small t values. Considering Top-$t = 1$, the performance of the baseline SMW algorithm is about 60.8%. Using other more sophisticated annotation algorithms, much better results are achieved, e.g., MtBGS achieves 74.2% which is the best among all the other compared algorithm, except the proposed algorithm WLRLCC, which can further boost the performance to 77.5%. The promising result validates the effectiveness of the proposed WLRLCC algorithm for

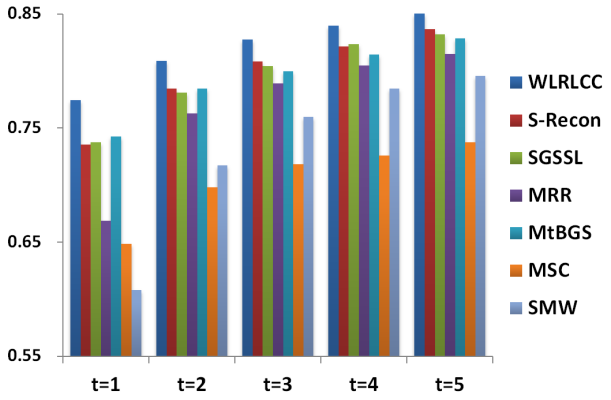


Figure 3: The comparison among WLRLCC and other algorithms on database GDB-040K.

improving the retrieval-based face annotation task. Secondly, similar to the observation obtained from Figure 2, more performance improvement is achieved by the WLRLCC algorithm for a small Top-t value, which is more critical for real applications. In practice, we mainly focus on a small t value since users usually would not be interested in a long list of annotated names. Thirdly, the annotation performance of MSC becomes worse when the t value for Top-t is too large. The MSC algorithm relies on a well-defined label matrix, which is used for graph construction and feature embedding. Although the label matrix Y is refined, it would still be noisy comparing to the manual labels.

	t=1	t=2	t=3	t=4	t=5
WLRLCC	0.7745 ± 0.012	0.8089 ± 0.010	0.8273 ± 0.010	0.8396 ± 0.009	0.8501 ± 0.009
S-Recon	0.7351 ± 0.010	0.7845 ± 0.010	0.8080 ± 0.010	0.8213 ± 0.008	0.8366 ± 0.007
SGSSL	0.7375 ± 0.013	0.7806 ± 0.009	0.8043 ± 0.009	0.8233 ± 0.009	0.8320 ± 0.009
MRR	0.6688 ± 0.014	0.7626 ± 0.009	0.7888 ± 0.009	0.8045 ± 0.009	0.8148 ± 0.011
MtBGS	0.7424 ± 0.013	0.7846 ± 0.013	0.7996 ± 0.011	0.8140 ± 0.011	0.8283 ± 0.010
MSC	0.6483 ± 0.013	0.6980 ± 0.013	0.7179 ± 0.013	0.7258 ± 0.011	0.7374 ± 0.010
SMW	0.6076 ± 0.015	0.7173 ± 0.009	0.7596 ± 0.008	0.7846 ± 0.007	0.7958 ± 0.007

Table 2: The comparison among WLRLCC and other algorithms on database GDB-040K.

5.6 Evaluation of Varied Database Sizes

In this experiment, we aim to examine the relationship between the annotation performance and the size of the retrieval database: GDB-040K, GDB-100K, GDB-200K, and GDB-600K. The experiment results are presented in Figure 1 and Table 3, where both the average performance and the standard deviation are presented. To be clear, we only show the annotation result with Top-$t = 1$, which is the most important case for the annotation task.

We can draw some observations from the results. First of all, similar to the previous observations, the WLRLCC algorithm consistently achieves the best annotation performance among all the compared algorithms for different database sizes. Second, it is clear that the annotation performance will decrease when increasing the size of the retrieval database, which is similar to the observation of the facial image retrieval in [37]. Note that there are two different ways for increasing the database sizes. One is to increase the number of facial images for each person, which is beneficial to the annotation task as more potential images are included, as shown in [33]; the other is to increase the number of unique persons, which could increase the difficulty of facial image retrieval and thus would degrade the annotation performance. The increase of database size in this experiment belongs to the second case. Third, the result also illustrates that the facial annotation problem remains a challenging problem, especially for large scale retrieval databases where accurate facial image retrieval could be a bottleneck for the annotation task, which is out of the scope of discussions in this paper.

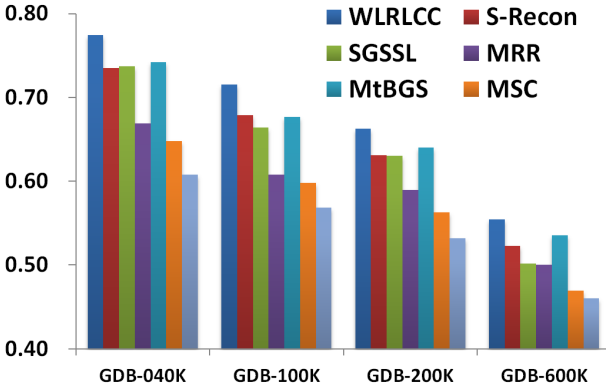


Figure 4: The comparison between WLRLCC and other algorithms on GDB-040K, GDB-100K,GDB-200K, and GDB-600K, with Top-t=1

5.7 Evaluation of Running Time Cost

In this section, we aim to evaluate the running time cost of the proposed WLRLCC algorithms, which consists of two major steps: (i) code learning and (ii) label learning. We adopt the FISTA [1] algorithm for solving the first problem, and implement the algorithm presented in Section 4.3.2 for the special QP problem in the second problem. To further reduce the running time, we also evaluate the performance of adopting PCA for dimension reduction over the original 512-dim GIST feature. We randomly select 1000 images from the retrieval database for PCA training. All the experiment results are presented in Table 4, where W-k denotes the new feature dimension is k after doing PCA projection.

Several observations can be drawn from these experiment results. First of all, the WLRLCC algorithm takes a bit longer time than the other algorithms, about 1.2 second for each query, as it has to code the feature and enhance the label together. Secondly, by using the PCA dimension reduction technique, the running time could be considerably reduced without introduce much performance degradation. For example, for the new 200-dim feature, the WLRLCC algorithm needs only about 0.63 second, but still achieves the best annotation performance, about 76.2%. Thirdly, when the dimension of the new feature space is small enough, the annotation performance of WLRLCC will decrease fast without more shrinkage in running time, which means there is a trade-off between the running time and the annotation performance. In this experiment, the suggested range of new feature dimension is about $[200, 300]$.

Besides the dimension reduction techniques, some other techniques could also been used for acceleration. In [39], Yang and Zhang propose a Gabor kernel based scheme to reduce the number

	WLRLCC	S-Recon	SGSSL	MRR	MtBGS	MSC	SMW
GDB-040K	0.7745±0.012	0.7351±0.010	0.7375±0.013	0.6688±0.014	0.7424±0.013	0.6483±0.013	0.6076±0.015
GDB-100K	0.7156±0.011	0.6789±0.009	0.6644±0.012	0.6080±0.010	0.6766±0.011	0.5981±0.012	0.5689±0.011
GDB-200K	0.6625±0.012	0.6314±0.008	0.6301±0.011	0.5898±0.008	0.6405±0.010	0.5628±0.012	0.5323±0.013
GDB-600K	0.5544±0.010	0.5228±0.007	0.5015±0.011	0.5005±0.008	0.5359±0.009	0.4694±0.010	0.4600±0.013

Table 3: Comparison between WLRLCC and other algorithms on GDB-040K, GDB-100K, GDB-200K, and GDB-600K (top-t=1).

	S-Recon	SGSSL	MRR	MtBGS	MSC
Time (s)	0.023 ± 0.007	0.971 ± 0.023	0.175 ± 0.016	0.414 ± 0.103	0.104 ± 0.191
Hit Rate	0.7351 ± 0.010	0.7375 ± 0.013	0.6688 ± 0.014	0.7424 ± 0.013	0.6483 ± 0.013
	WLRLCC	W-400	W-300	W-200	W-100
Time (s)	1.205 ± 0.041	0.959 ± 0.017	0.773 ± 0.028	0.63 ± 0.020	0.524 ± 0.030
Hit Rate	0.7745 ± 0.012	0.7639 ± 0.013	0.7619 ± 0.134	0.7615 ± 0.011	0.7403 ± 0.116

Table 4: The running time comparison of different algorithms and the annotation performance of different dimension reduction from the original 512-dim GIST feature(Top-t=1). W-k means the new feature dimension is k.

of atoms in dictionary, which could greatly reduces the computational cost for coding. In order to solve the large scale problem efficiently, some approximated algorithms are also proposed for sparse coding problem (e.g., LISTA [12], CoD [18], and so on.). These algorithms could also be adopted into our problem for running time reduction, which will be included in our further works.

5.8 Evaluation of Parameter Sensitivity

For the proposed WLRLCC algorithm, the best parameters λ_1 and λ_3 are found by a grid search on the GDB-040K database, which are also adopted in the experiments on other databases. Figure 5 shows the grid search results, where the ranges for λ_1, λ_3 are $\{0.005, 0.01, 0.1, 1, 3, 5\}$. From the figure, we observe that the performance of WLRLCC tends to be stable in some region $\lambda_1 \in [0.005, 0.1]$ and $\lambda_3 \in [0.1, 3]$. We thus choose $\lambda_1 = 0.01$ and $\lambda_3 = 1$ in our experiments, which are also adopted in the other three experiments on varied-scale databases. As shown in Figure 4, the encouraging performance of WLRLCC indicates that it is generally robust in the parameter setting for similar experiments.

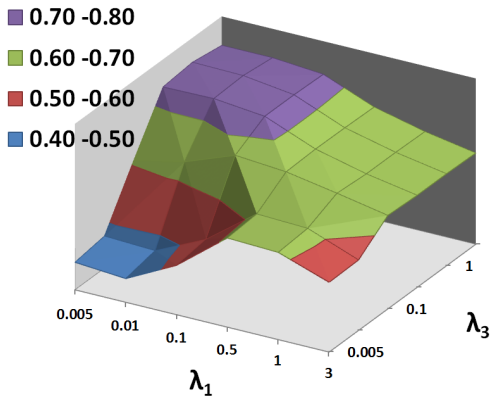


Figure 5: The grid search result of WLRLCC on GDB-040K.

6. DISCUSSIONS

There are two main assumptions for the proposed WLRLCC algorithm: (i) Although the initial weak label matrix is noisy and incomplete, the majority of its information remains beneficial; (ii) Two similar faces in the feature space share the same (close) name label. Typically, a generic image annotation problem also satisfies these assumptions. Our algorithm thus can be directly applied to a generic image annotation problem with a larger ε value for the multi-label task. On the other hand, the WLRLCC algorithm is usually limited by the discriminative ability of features, it would achieve a better performance on a special type of image annotation than generic image annotation problem.

The proposed retrieval-based face annotation scheme can be applied to many real-world applications. For example, it can be used to examine if a user uploads a celebrity's photo as his/her own avatar in a real-name social network. It can also be used for personal photo management in photo sharing web sites, where millions of unlabeled images are uploaded everyday.

For future work, we will address the issues of developing more efficient algorithms for the WLRLCC algorithm, including approximate sparse coding algorithms. In addition, instead of handling both the feature coding and label refinement steps in online manners, another way is to solve these problems over the whole retrieval database in an off-line manner. Future work will examine how to combine these two schemes effectively.

7. CONCLUSION

This paper investigated the retrieval-based face annotation problem and presented a promising framework to attack this challenge by mining massive weakly labeled facial images freely available on WWW. To improve the annotation performance, a novel Weak Label Regularized Local Coordinate Coding (WLRLCC) algorithm was proposed, which effectively exploits the principles of both local coordinate coding and graph-based weak label regularization. Using the achieved representative local coordinate coding and enhanced label matrix, a sparse reconstruction scheme is proposed for face name annotation. We conducted extensive experiments and found that the proposed WLRLCC algorithm achieved encouraging results on a large-scale web facial image testbed.

Acknowledgements

This work was supported by the Singapore National Research Foundation Interactive Digital Media R&D Program, under research grant NRF2008IDM-IDM004-006. Jianke Zhu was supported by Fundamental Research Funds for the Central Universities.

8. REFERENCES

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2:183–202, March 2009.

[2] T. L. Berg, A. C. Berg, J. Edwards, and D. Forsyth. Who's in the picture. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *NIPS*, Cambridge, MA, 2005. MIT Press.

[3] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. G. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In *CVPR*, pages 848–854, 2004.

[4] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *CVPR*, 2010.

[5] G. Carneiro, A. B. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Tran. PAMI*, pages 394–410, 2006.

[6] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, MA, 2006.

[7] J. Y. Choi, W. D. Neve, K. N. Plataniotis, and Y. M. Ro. Collaborative face recognition for improved face annotation in personal photo collections shared on online social networks. *IEEE Transactions on Multimedia*, 13, 2011.

[8] J. Cui, F. Wen, R. Xiao, Y. Tian, and X. Tang. Easyalbum: an interactive photo annotation system based on face clustering and re-ranking. In *CHI*, pages 367–376, 2007.

[9] W. Dong, Z. Wang, W. Josephson, M. Charikar, and K. Li. Modeling lsh for performance tuning. In *CIKM*, 2008.

[10] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.

[11] J. Fan, Y. Gao, and H. Luo. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *ACM Multimedia*, pages 540–547, 2004.

[12] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In J. Fürnkranz and T. Joachims, editors, *ICML*, pages 399–406. Omnipress, 2010.

[13] A. Hanbury. A survey of methods for image annotation. *J. Vis. Lang. Comput.*, 19:617–627, October 2008.

[14] E. Hjelmås and B. K. Low. Face detection: A survey. *In CVIU*, 83:236–274, 2001.

[15] P. O. Hoyer. Non-negative sparse coding. *CoRR*, cs.NE/0202009, 2002.

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[17] D.-D. Le and S. Satoh. Unsupervised face annotation by mining the web. In *ICDM*, pages 383–392, 2008.

[18] Y. Li and S. Osher. Coordinate descent optimization for l1 minimization with application to compressed sensing; a greedy algorithm. *Inverse Problems and Imaging*, 2009.

[19] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.

[20] J. Liu and J. Ye. Efficient euclidean projections in linear time. In *ICML*, pages 657–664, Montreal, Quebec, Canada, 2009.

[21] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *ACCV, 2010.*, June 2008.

[22] D. Ozkan and P. Duygulu. A graph based approach for naming faces in news photos. In *CVPR*, 2006.

[23] X. Rui, M. Li, Z. Li, W.-Y. Ma, and N. Yu. Bipartite graph reinforcement model for web image annotation. In *ACM Multimedia*, pages 585–594, Augsburg, Germany, 2007.

[24] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1-3):157–173, 2008.

[25] S. Satoh, Y. Nakamura, and T. Kanade. Name-it: Naming and detecting faces in news videos. *IEEE MultiMedia*, 6(1):22–35, 1999.

[26] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29:300–312, 2007.

[27] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Tran. Pattern Anal. Mach. Intell.*, 22(12): 1349–1380, 2000.

[28] Y.-Y. Sun, Y. Zhang, and Z.-H. Zhou. Multi-label learning with weak label. In *AAAI*. 2010.

[29] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain. Image annotation by knn-sparse graph-based label propagation over noisily tagged web images. *ACM Trans. Intell. Syst. Technol.*, 2:14:1–14:15, February 2011.

[30] Y. Tian, W. Liu, R. Xiao, F. Wen, and X. Tang. A face annotation framework with partial clustering and interactive labeling. In *CVPR*, 2007.

[31] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Image annotation refinement using random walk with restarts. In *ACM Multimedia*, pages 647–650, 2006.

[32] C. Wang, S. Yan, L. Zhang, and H.-J. Zhang. Multi-label sparse coding for automatic image annotation. In *CVPR*, 0:1643–1650, 2009.

[33] D. Wang, S.C.H. Hoi, and Y. He. Mining weakly labeled web facial images for search-based face annotation. In *SIGIR*, 2011.

[34] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. Annosearch: Image auto-annotation by search. In *CVPR*, 2006.

[35] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2), 2009.

[36] F. Wu, Y. Han, Q. Tian, and Y. Zhuang. Multi-label boosting for image annotation by structural grouping sparsity. In A. D. Bimbo, S.-F. Chang, and A. W. M. Smeulders, editors, *ACM Multimedia*, pages 15–24. ACM, 2010.

[37] Z. Wu, Q. Ke, J. Sun, and H.-Y. Shum. Scalable face image retrieval with identity-based quantization and multi-reference re-ranking. In *CVPR*, pages 3469–3476, 2010.

[38] J. Yang and A. G. Hauptmann. Naming every individual in news video monologues. In *ACM Multimedia Conference*, pages 580–587, New York, NY, USA, 2004.

[39] M. Yang and L. Zhang. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In *ECCV*, pages 448–461, 2010.

[40] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *NIPS*, pages 2259–2267, 2009.

[41] L. Zhang, L. Chen, M. Li, and H. Zhang. Automated annotation of human faces in family albums. In *ACM Multimedia*, 2003.

[42] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.

[43] J. Zhu, S.C.H. Hoi, and L.V. Gool. Unsupervised face alignment by robust nonrigid mapping. In *ICCV*, 2009.

[44] J. Zhu, S.C.H. Hoi, and M.R. Lyu. Face Annotation Using Transductive Kernel Fisher Discriminant. *IEEE Transactions on Multimedia.*, 10(1):86–96, 2008.

[45] X. Zhu, Z. Ghahramani, and J.D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *The Twentieth International Conference on Machine Learning (ICML-2003)*, pages 912–919, 2003.