Dual Discriminative Local Coding for Tissue Aging Analysis

Yang Song\textsuperscript{a,*}, Qing Li\textsuperscript{a}, Fan Zhang\textsuperscript{a}, Heng Huang\textsuperscript{b}, Dagan Feng\textsuperscript{a,c}, Yue Wang\textsuperscript{d}, Mei Chen\textsuperscript{e,f}, Weidong Cai\textsuperscript{a}

\textsuperscript{a}School of Information Technologies, University of Sydney, Australia
\textsuperscript{b}Department of Computer Science and Engineering, University of Texas at Arlington, USA
\textsuperscript{c}Med-X Research Institute, Shanghai Jiaotong University, China
\textsuperscript{d}Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, USA
\textsuperscript{e}Computer Engineering Department, University of Albany State University of New York, USA
\textsuperscript{f}Robotics Institute, Carnegie Mellon University, USA

Abstract

In aging research, morphological age of tissue helps to characterize the effects of aging on different individuals. While currently manual evaluations are used to estimate morphological ages under microscopy, such operation is difficult and subjective due to the complex visual characteristics of tissue images. In this paper, we propose an automated method to quantify morphological ages of tissues from microscopy images. We design a new sparse representation method, namely dual discriminative local coding (DDLC), that classifies the tissue images into different chronological ages. DDLC incorporates discriminative distance learning and dual-level local coding into the basis model of locality-constrained linear coding thus achieves higher discriminative capability. The morphological age is then computed based on the classification scores. We conducted our study using the publicly available terminal bulb aging database that has been commonly used in existing microscopy imaging research. To represent these images, we also design a highly descriptive descriptor that combines several complementary texture features extracted at two scales. Experimental results show that our method

\textsuperscript{*}Corresponding author.
Email address: yson1723@uni.sydney.edu.au (Yang Song)
achieves significant improvement in age classification when compared to the existing approaches and other popular classifiers. We also present promising results in quantification of morphological ages.

**Keywords:** microscopy image, tissue age, sparse representation

### 1. Introduction

With the longer life expectancy, aging-related health issues such as cancer and neurodegenerative diseases are becoming more prevalent. Improving the quality of life at advanced age has become a critical problem in public health, and understanding the causal processes of aging is essential to facilitate this development. However, currently there is no established theory of aging. While some researchers suggest aging is predetermined by genes, others believe it is the result of many lifelong influences (Sergiev et al., 2015). Aging studies have thus taken a multidisciplinary approach involving gene expression analysis, proteomics and metabolomics, pathway analysis, and image analysis (Wieser et al., 2011).

Image-based aging studies have focused on morphological age estimation from microscopy images (Wieser et al., 2011). Morphological age refers to the quantitative measure of age based on tissue morphology that can be visualized from microscopy images. It can represent the actual effect of aging on the particular individual, and individuals of the same chronological age can have different morphological ages. This information is important in aging research to discover the various factors causing different effects of aging on different people.

To quantify the morphological age of tissues, manual approaches have been conducted (Garigan et al., 2002; Herndon et al., 2002; Helfrich et al., 2007). These studies used manual scoring of morphological ages as the basis for further analysis and made interesting discoveries about the aging process. However, these approaches used different ways of scoring, without a standardization of the scoring process. Manual scoring can also be difficult to reproduce since it is hard to describe the visual criteria used for scoring. Especially if a more detailed quantification is expected rather than just differentiating between a few age categories, the manual approach would be subjective and unreliable. Techniques that can facilitate automatic quantification of tissue morphological age are highly desirable. In this paper, we present our approach to this problem.
1.1. Related Work

Computer-aided approaches have been proposed for automated classification of tissue ages (Shamir et al., 2008b; Johnston et al., 2008; Shamir et al., 2009; Meng and Shyu, 2013; Coelho et al., 2013; Zhou et al., 2013; Pham, 2014). These approaches typically work by first extracting image features from various aspects then performing classification to separate the images into different age groups. For image feature description, texture features have mainly been used, including the traditional Haralick, Tamura and wavelet features (Shamir et al., 2008b; Johnston et al., 2008; Shamir et al., 2009; Zhou et al., 2013), more recent speeded-up robust features (SURF) (Coelho et al., 2013), and customized entropy features (Pham, 2014). Classification is then performed using various algorithms, such as the nearest neighbor classifier (Shamir et al., 2008b; Pham, 2014), Fisher discriminant analysis (Johnston et al., 2008), and support vector machine (SVM) (Coelho et al., 2013; Zhou et al., 2013). Evaluation is then performed by comparing the derived age groups with the annotated chronological ages.

With the large variety of techniques, the age classification performance is however quite low with most of the studies reporting around 50% accuracy. Figure 1 shows the example images from the terminal bulb aging database (Shamir et al., 2008a), which is commonly used in the aforementioned studies. It can be seen that there is a general trend of structural deterioration.
of muscle tissues with aging. However, it is hard to distinguish the degree of deterioration at different ages, and images of the same chronological age show large variations of visual characteristics. In addition, these approaches provide classification of chronological ages only, and do not support quantification of morphological ages. This could be due to the unavailability of ground truth for morphological ages, which is still a subjective measure in current aging research.

In a broader scope, tissue aging analysis is closely related to microscopy image classification, since they have the same overall framework that comprises feature extraction and classification. Depending on the focus of method design, existing studies in microscopy image classification can be categorized into two groups. The first group focuses on feature extraction, in which customized features are designed (Su et al., 2012; Sparks and Madabhushi, 2013; Peter et al., 2015; Xu et al., 2015; Jiang et al., 2015; Barker et al., 2016) or automated feature learning is conducted (Zhou et al., 2014; Otalora et al., 2015; BenTaieb et al., 2015; Wang et al., 2015). The second group focuses on the classifier design while standard and simple feature descriptors are used. In particular, multiple instance learning (Xu et al., 2014; Kandemir et al., 2014; Li et al., 2015) and sparse representation (Srinivas et al., 2014; Vu et al., 2015; Taalimi et al., 2015; Su et al., 2015) have been the major works in classifier design for microscopy image classification.

In this work we focus on classifier design and propose a sparse representation based method. Here we present a detailed review on sparse representation and its application in biomedical image analysis. Sparse representation performs classification based on the quality of reconstruction for the classification tasks (Wright et al., 2010). Typically the test data is sparsely reconstructed using reference dictionaries of different classes, and then classified to the class corresponding to the best reconstruction. Unlike SVM and other discriminative or generative classifiers, sparse representation is a non-parametric model and does not require training to model the separation between classes. Its data-driven approach is especially suitable when there is limited amount of training data to effectively exploit the discriminative structure in the feature space. Sparse representation has thus been a popular classifier for biomedical imaging applications (Xu et al., 2013; Weiss et al., 2013; Song et al., 2013; Srinivas et al., 2014; Wang et al., 2014).

On the other hand, the classification performance of sparse representation is entirely dependent on the quality of sparse reconstruction, and the classification objective is not associated with the reconstruction process. Since
sparse reconstruction is optimized towards achieving good reconstruction of the test data given any reference dictionary, a better reconstruction could be obtained for the wrong class, leading to misclassification. Several ways have been proposed to improve the basic sparse representation model. For example, dictionary learning is used to generate a more discriminative reference dictionary from the image patches or feature vectors (Liu et al., 2011; Tong et al., 2013; Vu et al., 2015; Taalimi et al., 2015). Such methods are suitable for problems with large number of training data and small feature dimension. Another way is to restrict the selection of reference items that are used to reconstruct the test data. In locality-constrained linear coding (LLC) (Wang et al., 2010), the nearest neighbors of the test data are first identified and then linearly combined to obtain the sparse representation. LLC is highly efficient with its analytical solution and has been applied in various biomedical imaging studies (Zhang et al., 2013; Xing and Yang, 2013; Wu et al., 2014; Song et al., 2015b; Su et al., 2015). In addition, ensemble learning has been integrated with sparse representation, in which the sparse reconstruction is performed using subsets of reference dictionaries and the classification decisions are fused with boosting (Huang et al., 2014; Song et al., 2014), large margin optimization (Song et al., 2015a), or customized fusion weights (Song et al., 2015b). These approaches impose an additional layer of discrimination based on the reconstruction outputs and can effectively enhance the classification performance.

1.2. Our Contribution

In this work, we present an automated method for quantifying the morphological age of tissues from microscopy images. Since currently there is no established rules to compute the morphological ages and automated approaches only support classification of chronological ages, our design is exploratory and we hypothesize that the tissue morphological age should be closely correlated with its chronological age. We thus propose to derive the morphological age by first classifying the chronological age of the tissue. The classification outputs give the degrees of correlation between the test image and various chronological ages. Then based on these correlation measures, the morphological age is computed. In summary, our method comprises three components: image feature extraction, chronological age classification, and morphological age quantification. Figure 2 illustrates our method design.

Our main methodological contribution is the design of a new sparse representation model, namely the dual discriminative local coding (DDLC)
method, which is used for the chronological age classification component. Briefly, we choose the LLC model as our base classifier, considering its computational efficiency and improved classification performance with the locality constraints. However, the discriminative capability of LLC can still be limited, because LLC does not involve learning-based optimization for classification and it derives the locality constraints simply based on Euclidean distances. Our DDLC method is designed to address these issues with discriminative distance computation and dual-level local coding. Our second contribution is that we have designed a new way of texture representation for the terminal bulb images. By combining multiple types of texture descriptors extracted at two levels, we obtain a highly descriptive image descriptor. Finally, we propose a method that quantifies the morphological ages based on the classification outputs of chronological ages. Evaluation is performed on the publicly available terminal bulb aging database (Shamir et al., 2008a), which is commonly used in studies for automated tissue aging analysis.

A preliminary version of this work has been published in the conference paper (Song et al., 2016). Different from the two-level LLC model proposed in the preliminary study, in this work, we have further incorporated discriminative learning into the sparse representation model and developed the new DDLC method. In addition, more thorough evaluation of the quantification results has been conducted. This paper provides substantially more detailed description of the problem domain and algorithm design.
2. Dual Discriminative Local Coding

We first describe the preliminaries of the LLC algorithm and an LLC-based classification method. We then elaborate on the design motivation of our DDLC method, followed by presenting our design of dual-level local coding and discriminative distance computation.

2.1. Background

LLC is originally used as a descriptor coding algorithm that converts a feature descriptor into a sparse coding representation based on reference dictionaries. The essential difference between LLC and the other well-known sparse coding algorithms is the inclusion of locality constraint, which ensures that the sparse reconstruction is obtained from reference vectors that are similar to the input descriptor. LLC is also more computationally efficient with an analytical solution.

Formally, consider that each image is represented by an $H$-dimensional feature descriptor, and the training set contains $N$ images. Let $x \in \mathbb{R}^{H \times 1}$ denote the feature descriptor of a test image $I$, and $B \in \mathbb{R}^{H \times N}$ denote the reference dictionary constructed from the training set. The LLC algorithm is formulated as follows:

$$\min_{c_x} \|x - Bc_x\|^2 + \lambda \|d_x \odot c_x\|^2 \quad \text{s.t.} \quad \mathbf{1}^T c_x = 1, \quad \|c_x\|_0 = K \quad (1)$$

where $d_x \in \mathbb{R}^{N \times 1}$ contains the Euclidean distance between $x$ and each reference vector in $B$, $\lambda$ is a constant, and $c_x \in \mathbb{R}^{N \times 1}$ is a $K$-sparse reconstruction coefficient. To obtain $c_x$, $K$ reference vectors that are most similar to $x$ are first identified from $B$. Based on these $K$ nearest neighbors, $c_x$ is then derived analytically (see details in Wang et al., 2010). Subsequently, $c_x$ is the LLC code of input $x$ and is typically used with SVM for classification. The image $I$ can also be divided into sub-regions and the region-level codes can be pooled together to obtain the image-level code.

In our study, we apply LLC as a sparse representation classifier. This is performed by first partitioning the reference dictionary $B$ into $L$ dictionaries, with $L$ denoting the number of image classes. Each dictionary $B_l : l = 1, ..., L$ contains the reference data of class $l$. For a test input $x$, LLC is then applied to obtain $c_{x,l}$, the coding of $x$ respective to each class:

$$\min_{c_{x,l}} \|x - B_{l}c_{x,l}\|^2 + \lambda \|d_{x,l} \odot c_{x,l}\|^2 \quad \text{s.t.} \quad \mathbf{1}^T c_{x,l} = 1, \quad \|c_{x,l}\|_0 = K \quad (2)$$
Finally, the distance between $x$ and its reconstruction from each dictionary is computed, and $x$ is classified based on the smallest distance:

$$\arg\min_l \|x - B_l c_{x,l}\|^2$$

(3)

2.2. Design Motivation

While we have obtained good results using the LLC classifier in our previous study (Song et al., 2015b), we observe two issues in the model that could reduce the classification performance. First, the LLC classifier does not involve discriminative learning. The classification performance is dependent on the reconstruction performance using the reference dictionaries of the various classes, but there is no explicit constraint to encourage better reconstruction from the dictionary of the correct class. Second, the $K$ nearest neighbors are determined using Euclidean distances between the input $x$ and the reference vectors. Since the feature vectors normally provide limited discriminative power, the Euclidean distance between vectors would not reflect the visual similarity well and the $K$ nearest neighbors could be misleading.

Our DDLC method is designed to address these issues with the following algorithms. First, we incorporate discriminative distance computation into the LLC model (Section 2.3) so that the distance between the input $x$ and its reconstruction from the right class would be less than the distance between $x$ and its reconstruction from the wrong class. This approach helps to incorporate supervised learning into the distance-based classification model. Second, we propose to a dual-level local coding (Section 2.4) that identifies two sets of nearest neighbors at the global level and local cluster level, and fuses the classification decisions from the two levels to obtain the final classification. The inclusion of nearest neighbors at the local cluster level helps to explore the feature similarities in an unsupervised manner.

2.3. Discriminative Distance Computation

Our design of discriminative distance computation is to learn a dimension reduction matrix, so that in the dimension-reduced feature space, we obtain better reconstruction from the right class therefore more accurate classification.

Specifically, assume that we have a linear projection matrix $W \in \mathbb{R}^{H' \times H}$, $H' < H$. The feature descriptor $x$ is then projected into vector $Wx$ of dimension $H'$, and its reconstruction $B_l c_{x,l}$ is projected into $WB_l c_{x,l}$ as well.
In this lower dimensional space, we expect the distance between $x$ and its reconstruction from the correct class $l$ to be less than a threshold $b \in \mathbb{R}$:

$$
\|WX - WB_l c_{x,l}\|^2 < b
$$

and the distance between $x$ and its reconstruction from the wrong class to be greater than $b$. As a result, $x$ would be correctly classified. Figure 3 illustrates the effect of this process.

To learn $W$ and $b$, we formulate our learning objective by imposing a distance margin of at least 1:

$$
y_{xl}(b - \|WX - WB_l c_{x,l}\|^2) > 1, \ \forall l = 1, ..., L
$$

where $y_{xl} = 1$ if $x$ is of class $l$, and $y_{xl} = -1$ otherwise. With a training set of $N$ descriptors $\{x_n : n = 1, ..., N\}$, the objective function can then be rewritten as:

$$
\arg\min_{W,b} \sum_{n,l} \max[1 - y_{nl}(b - \|WX_n - WB_l c_{n,l}\|^2), 0]
$$
Here \( c_{n,l} \) denotes the coding of \( x_n \) with dictionary \( B_l \), and is derived using Eq. (2).

To solve Eq. (6), \( W \) is first initialized by performing principal component analysis (PCA) on the entire dataset with \( H' \) set to half of the number of images. A stochastic sub-gradient method (Algorithm 1) is then applied to optimize \( W \) and \( b \). Note that \( \gamma \) and \( \beta \) are constant learning rates and default to 0.25 and 1. Subsequently, the input \( x \) can be classified by the following:

\[
\arg\min_l \| Wx - WB_l c_{x,l} \|^2
\]  

(7)

**Algorithm 1: Solving Eq. (6).**

**Data:** Training data \( \{x_n : n = 1, \ldots, N\} \), and their reconstructions from each dictionary \( \{B_l c_{n,l} : n = 1, \ldots, N; l = 1, \ldots, L\} \).

**Result:** Linear projection matrix \( W \).

Initialize \( W_0 \) using PCA on the training data;

Initialize \( b_0 \) as the threshold that can satisfy

\( y_{nl}(b_0 - \| W_0 x_n - W_0 B_l c_{n,l} \|^2) > 1 \) for most training data;

Initialize iteration number \( t = 0 \);

repeat

Sample a training data \( \{x_n, B_l c_{n,l}\} \);

if \( y_{nl}(b_t - \| W_t x_n - W_t B_l c_{n,l} \|^2) \leq 1 \) then

\( \Delta_{nl} = (x_n - B_l c_{n,l})(x_n - B_l c_{n,l})^T \);

\( W_{t+1} = W_t - \gamma y_{nl} W_t \Delta_{nl} \);

\( b_{t+1} = b_t + \beta y_{nl} \);

end

\( t = t + 1 \);

until convergence or maximum number of iterations is reached;

We would like to note that while our way of solving Eq. (6) is similar to the metric learning method (Simonyan et al., 2013), our formulation of the learning objective is conceptually different. Specifically, our distance computation is between the input descriptor and its reconstructions rather than between pairs of input descriptors, and is adapted to enhance the discriminative capability of the sparse representation classification.
2.4. Dual-level Local Coding

Our design of dual-level local coding is to reconstruct the input descriptor $x$ using (i) the global reference dictionary $B$ consisting of all training data, and (ii) a subset of $B$ constructed from the cluster containing $x$. The cluster-level coding introduces an additional layer of similarity computation with unsupervised clustering and restricts the selection of $K$ nearest neighbors (used for reconstructing $x$) to the cluster.

Since the global-level coding is the same as the original LLC model (Section 2.1), we only describe how to obtain the cluster-level coding here. To do this, we first cluster the dataset into $M$ clusters using the locality-constrained subspace clustering (LSC) method (Song et al., 2015c). LSC is an unsupervised clustering algorithm that is based on spectral clustering. Briefly, given the dataset of $Q$ images (including both training and testing data), LSC works by first generating an affinity matrix $A \in \mathbb{R}^{Q \times Q}$ using sparse coding to represent the similarities between images, then the affinity matrix $A$ is used in spectral clustering to obtain $M$ clusters. LSC is related to the sparse sub-
space clustering (SSC) algorithm (Elhamifar and Vidal, 2009), but is more efficient than SSC by replacing the L1 regularized sparse coding with LLC; and it has been shown that LSC provides better clustering performance than k-means and spectral clustering (Song et al., 2015c). In this clustering process, the class label information is unused.

Next, assume that the input descriptor $x$ is partitioned into cluster $m$. The training data belonging to cluster $m$ are then extracted to form a cluster-level reference dictionary $B^m = \{B^m_l : l = 1, ..., L\}$. The cluster-level coding of $x$ respective to class $l$ is then computed as $c_{x,l}^m$:

$$\min_{c_{x,l}^m} \|x - B^m_l c_{x,l}^m\|^2 + \lambda \|d_{x,l}^m \odot c_{x,l}^m\|^2 \quad \text{s.t.} \quad 1^T c_{x,l}^m = 1, \|c_{x,l}^m\|_0 = \lfloor K/M \rfloor \quad (8)$$

In addition, the linear projection matrix $W'$ is also learned at the cluster-level, so that we expect to obtain better cluster-level reconstruction with the help from the discriminative learning. Specifically, we formulate the following objective function:

$$\arg\min_{W', b} \sum_{n,l} \max[1 - y_{nl}(b' - \|W' x_n - W' B^m_l c_{x,l}^m\|^2), 0] \quad (9)$$

where $W'$ and $b'$ denote the cluster-level linear projection matrix and threshold, $B^m_l$ indicates the cluster-level reference dictionary of the training input $x_n$, and $c_{x,l}^m$ is the cluster-level coding of $x_n$. Algorithm 1 is adapted to solve this equation. Note that in order to maximize the number of training data, here a single $W'$ is learned for all the clusters rather than learning one matrix for each cluster.

Finally, by fusing results from both global- and cluster-level coding, we obtain the classification of $x$ by:

$$\arg\min_l \|W x - WB_l c_{x,l}\|^2 + \alpha \|W' x - W' B^m_l c_{x,l}^m\|^2 \quad (10)$$

where $\alpha$ is a constant. This is the final output of our DDLC method. Figure 4 illustrates the design of dual-level local coding.

3. Tissue Aging Analysis

In this section, we describe the dataset used for tissue aging studies, our design of image feature descriptors, and the way of using our dual discriminative local coding for morphological age quantification. The experimental setup is also described.
Table 1: Summary of the terminal bulb aging dataset.

<table>
<thead>
<tr>
<th>Age (days)</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td># images</td>
<td>106</td>
<td>218</td>
<td>159</td>
<td>176</td>
<td>195</td>
<td>62</td>
<td>54</td>
</tr>
</tbody>
</table>

3.1. Image Dataset

The publicly available terminal bulb aging dataset from the IICBU 2008 database (Shamir et al., 2008a) is used in this study. The database contains 970 differential interference contrast (DIC) microscopy images captured from the pharynx terminal bulb of *C. elegans*. *C. elegans* have a short lifespan of 2–3 weeks, and the morphological changes of tissues during aging can be easily observed using a microscope. These properties make *C. elegans* a valuable experimental organism for the study of aging. Each image is of 300×300 pixels, and annotated with a chronological age (0, 2, 4, 6, 8, 10, or 12 days). Table 1 lists the number of images belonging to each chronological age. Sample images are shown in Figure 1.

3.2. Texture Feature Extraction

We use four types of texture descriptors to represent the image feature: local binary patterns (LBP) (Ojala et al., 2002), histogram of oriented gradients (HOG) (Felzenszwalb et al., 2010), GIST (Oliva and Torralba, 2001) and census transform histogram (CENTRIST) (Wu and Rehg, 2011). The selection of these features is based on three considerations: (i) we tested features that have been traditionally used in microscopic image analysis (e.g. the gray-level co-occurrence matrix, wavelet and curvelet transforms, and LBP); (ii) we incorporated features that are commonly effective in computer vision such as HOG, GIST, CENTRIST, and improved Fisher vector (IFV) (Perronnin et al., 2010); and (iii) we experimented with convolutional neural network (CNN) including the 4096-dimensional features extracted from the last fully connected layer of the VGG-VD model pretrained on ImageNet (Simonyan and Zisserman, 2015; Cimpoi et al., 2015), and a model trained on our dataset using a simplified AlexNet architecture. We found the combination of the selected four features provided the best performance.

The features are extracted at two scales. At the coarse scale, the descriptors are extracted for the whole image. In particular, the descriptors are computed following the standard settings, and the image is divided into
16 cells for concatenating the cell-level LBP / HOG / GIST descriptors into image-level descriptors. This results in LBP, HOG, GIST, and CENTRIST descriptors of $16 \times 58$, $16 \times 31$, $16 \times 32$, and 256 dimensions, respectively.

At the fine scale, the 16-cell LBP and HOG descriptors are extracted for the middle region of image. The middle region is centered at the image center, and its width / length is half of the image width / length. We perform this fine-scale feature extraction because we observe that more visual differences can be seen from the middle region than the border areas when comparing images of different ages. In addition, our experimental analysis shows that GIST and CENTRIST features from the middle region would introduce more noise and should not be included.

The image descriptor is the concatenation of all individual descriptors, and is of $16 \times (58 + 31 + 32) + 256 + 16 \times (58 + 31) = 3616$ dimensions. Note that each individual descriptor is L2 normalized before concatenation.

3.3. Morphological Age Quantification

To quantify the morphological age of a terminal bulb sample, we design a two-step approach: first, we apply the DDLC method (Section 2) to obtain the chronological age class of the image; then, the morphological age is computed based on the classification outputs. Our design is mainly guided by the following considerations. First, the terminal bulb aging dataset provides only the chronological age of each image. The ground truth of morphological ages is not available and currently there is no standard rules to compute the morphological age quantitatively from image inputs. This means that we could only use the chronological ages as labels in learning-based approaches. Second, the previous studies (Garigan et al., 2002; Johnston et al., 2008; Shamir et al., 2009) show that there are distinct morphological states in the aging process and these states have good correlation with the chronological ages. We thus suggest that the morphological age of an image can be derived based on the degree of correlation between the image and various chronological ages.

In the first step, the chronological age of a test image is obtained by classifying it into one of seven classes (Table 1) using the DDLC method, i.e. Eq. (10). Here, seven reference dictionaries $\{B_l : l = 1, \ldots, L = 7\}$ are constructed from the training data, which are also used to learn the projection matrices $W$ and $W'$. Since the number of training data is small relative to the number of variables in the projection matrices, over-fitting would be an issue. To limit the effect of over-fitting, the number of iterations performed during
the optimization process is set to $10^5$ for an early stopping. In addition, $M = 3$ clusters are used for the cluster-level coding, and the constant factor $\alpha$ is set to 0.5. When generating the codings $c_{x,l}$ and $c_{x,l}^m$, the constant $K$, which is the number of nearest neighbors used in Eqs. (2) and (8), is set to 50. The parameter settings for $M$, $\alpha$, and $K$ are determined from our initial experimental studies with one split of five-fold cross validation.

In the second step, the morphological age is derived based on the classification outputs of chronological ages. We compute the score $s(x, l)$ of $x$ belonging to each class $l$ as:

$$s(x, l) = \exp\{\beta (1 - \frac{r(x, l)}{\sum_{l' = 1}^{L} r(x, l')})\}$$

where $r(x, l) = \|W x - WB_{l}c_{x,l}\|^2 + \alpha\|W'x - W'B_{l}^{m}c_{x,l}^{m}\|^2$ as defined in Eq. (10) and derived in the first step when classifying the chronological age, and $\beta > 1$ is a constant. A lower $r(x, l)$ leads to a higher $s(x, l)$, which means a higher correlation between $x$ and class $l$. The exp function and scaling constant $\beta$ (default to 10) are used to enlarge the score difference between different classes.

Assuming that the highest score is obtained for class $l$, we then generate a continuous label $v_x \in (1, L = 7)$ for $x$ by:

$$v_x = \frac{\sum_{l' = l_1}^{l_2} l' s(x, l')}{\sum_{l' = l_1}^{l_2} s(x, l')}$$

where $\delta \in \{1, 2, 3\}$ is a constant parameter and specifies the number of neighbors of $l$ used to compute the label. By default we set $\delta = 1$ so that only the adjacent neighbors of $l$ are included. Note that $\delta$ is maximally 3, which would include all $L = 7$ classes into the computation. Finally, the morphological age $a_x$ is a rescaling of $v_x$ to shift $v_x$ into the range of 0 to 12 days: $a_x = 2v_x - 2$, to align with the range of chronological ages of the dataset.

3.4. Experimental Setup

We conducted five-fold cross validation for evaluation. In particular, 4/5 of the images from each class were randomly selected for training and the rest were used testing. We repeated the cross validation for ten times with ten different splits of training and testing data. Following the existing studies (Shamir et al., 2008b; Zhou et al., 2013; Coelho et al., 2013), we used
Table 2: Comparison with existing studies.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
<th>Feature</th>
<th>Classifier</th>
<th>Setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>71.9±2.7</td>
<td>LBP, HOG, GIST, CENTRIST</td>
<td>DDLC</td>
<td>10 splits of 5-fold CV</td>
</tr>
<tr>
<td>Shamir et al. 2008b</td>
<td>49</td>
<td>Radon, Gabor, Tamura, edge, object, Zernike, etc</td>
<td>kNN</td>
<td>one split of training and testing sets</td>
</tr>
<tr>
<td>Zhou et al. 2013</td>
<td>51.1</td>
<td>Hu moments, Zernike, Hessian, wavelet, etc</td>
<td>SVM</td>
<td>5 splits of 5-fold CV</td>
</tr>
<tr>
<td>Coelho et al. 2013</td>
<td>44.6</td>
<td>SURF, Haralick, object, overlap, etc</td>
<td>SVM</td>
<td>10-fold CV</td>
</tr>
</tbody>
</table>

CV: cross validation.

classification accuracy as the evaluation metric. The mean and standard deviation of the chronological age classification results were measured and compared with existing results and other classifiers. In addition, to evaluate the quantification outputs of the morphological ages, we focused on qualitative analysis of the resultant distributions of morphological ages corresponding to different chronological ages. Our method was implemented in Matlab, running on a PC with Intel i7 CPU and 16 GB memory.

4. Results and Discussion

4.1. Chronological Age Classification

We first evaluate our method performance for classifying chronological ages, since existing approaches on automated tissue analysis mainly focus on such classification problems. Table 2 shows the classification results in comparison with the published results for the same terminal bulb database (Shamir et al., 2008b; Zhou et al., 2013; Coelho et al., 2013). We computed the mean and standard deviation of classification accuracies from the ten splits of five-fold cross validation, while the published results only reported the mean accuracy. It can be seen that our method outperformed the related work by a large extent.

Table 2 also summarizes the feature descriptors, classifiers, and experimental setups used in the various approaches. Our approach incorporated
more recent descriptors and we designed a new DDLC model, while the compared approaches used more traditional features and the standard SVM and $k$NN classifiers.

To analyze the effect of our proposed feature design and classification model, we evaluated the classification performance with F-SVM and F-$k$NN, which denote the approaches combining our feature descriptor with SVM and $k$NN classifiers, respectively. For SVM, we found that the linear kernel was more effective for this task compared to the other commonly used non-linear kernels. For $k$NN, we used the same number of neighbors as used in our DDLC model. Figure 5 shows the various classification accuracies. The improvement of F-SVM and F-$k$NN over (Shamir et al., 2008b; Zhou et al., 2013; Coelho et al., 2013) indicates that our feature design can better describe the terminal bulb images than the more traditional texture features. In addition, the advantage of our approach (F-DDLC) over F-SVM and F-$k$NN demonstrates that our DDLC model was more discriminative than SVM and $k$NN for the age classification task.

To further evaluate our DDLC method, we measured the classification accuracies using the following variations of DDLC: (i) the basic L0-regularized sparse representation (SR0); (ii) the basic L1-regularized sparse representation (SR1); (iii) sparse representation using the standard LLC; (iv) the LLC model with additional discriminative distance learning, which is denoted as
Figure 6: Average classification accuracy and standard deviation, comparing our DDLC method with the other variations based on sparse representation.

Dist; and (v) the LLC model with dual-level local coding, which is denoted as Dual. In addition, we tested using the Fisher discrimination dictionary learning (FDDL) (Yang et al., 2011) method as the classifier. FDDL is a popular benchmark used in general imaging for evaluating sparse representation techniques with discriminative enhancement. We also tested using the collaborative representation classification (CRC) with regularized least square (Zhang et al., 2011). CRC imposes a weaker sparsity constraint and focuses more on the collaborative representation or reconstruction by all training data. For all these compared approaches, our proposed feature descriptor was used, and the parameters were tuned using one split of five-fold cross validation as well.

As shown in Figure 6, our DDLC method achieved the highest classification accuracy. With the additional locality constraints, LLC outperformed SR0 and SR1 with about 5.7% and 2.4% improvement, and is much faster requiring about 0.03 s per image compared to 0.9 s and 0.12 s per image with SR0 and SR1. LLC is thus a good choice as our basis classifier. The advantages of Dist and Dual over LLC indicate that our discriminative distance learning and dual-level local coding can further improve the discriminative power of LLC. The performance improvement of DDLC over Dist and Dual demonstrates the benefit of integrating both components. While DDLC is slower than LLC with about 0.06 s per image (mainly due to the inclusion of
a second cluster-level coding), it provides 5.6% improvement in classification accuracy. Furthermore, we found that FDDL performed similar to SR0. This suggests that with the small size of training data relative to the feature dimension, the discriminative encoding of dictionary has limited effect, while in our DDLC model, the discriminative learning is more directly coupled with the classification objective. CRC provided the lowest performance, indicating that for this dataset sparse reconstruction could provide more accurate classification compared to collaborative reconstruction.

The benefit of Dist (discriminative distance learning) comes from the discriminative dimension reduction with linear projection, which is optimized to reduce the reconstruction error for the correct class. Figure 7 illustrates the effect of this linear projection, obtained during one run of cross validation. The training time required for the linear project matrix was about 4 minutes. The improvement provided by Dual (dual-level local coding) is attributed to the cluster-level coding. The clustering step effectively altered the similarity relationship between the test data and reference vectors. For example, in our experiment, 50 and 16 nearest neighbors were used as the reference vectors for global- and cluster-level coding. On average 12 of the reference vectors were common between the two levels of coding. The remaining differences in reference sets introduced diversity in reconstruction and led to the improved classification performance after fusing the global- and cluster-level coding.
Table 3: Confusion matrix of age classification.

<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>Prediction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Day 0</td>
</tr>
<tr>
<td>Day 0</td>
<td>98.1</td>
</tr>
<tr>
<td>Day 2</td>
<td>0.9</td>
</tr>
<tr>
<td>Day 4</td>
<td>0</td>
</tr>
<tr>
<td>Day 6</td>
<td>0</td>
</tr>
<tr>
<td>Day 8</td>
<td>0</td>
</tr>
<tr>
<td>Day 10</td>
<td>0</td>
</tr>
<tr>
<td>Day 12</td>
<td>0</td>
</tr>
</tbody>
</table>

In addition, we found that the choice of clustering algorithm was important, and clustering using LSC provided 2% improvement in classification accuracy compared to using \( k \)-means to generate the clusters. Slightly higher (0.6%) classification accuracy was obtained with \( M = 3 \) clusters compared to \( M = 2 \), but the performance degraded with larger \( M \). This suggests that the size of cluster cannot be too small, which would reduce the reconstruction capability of sparse representation. Also, we found that the parameter \( K \), which specified the number of nearest neighbors used in reconstruction, had a small influence on the classification results. By varying \( K \) between 20 and 100, the average classification accuracies were between 71.1% and 72.4%, while our adopted setting of \( K = 50 \) provided 71.9% accuracy.

We also evaluated if the performance improvement by DDLC over SVM, \( k \)NN, SR0, SR1, LLC, FDDL and CRC was statistically significant. One-tailed paired t-test was applied for this analysis. The null hypothesis was that DDLC provided the same classification results as the paired approach. We obtained p-value < \( 10^{-4} \) for all comparisons. These results thus indicate that our DDLC method achieved statistically significant improvement over the compared classifiers.

Finally, we analyzed the classification performance of the individual age classes. Table 3 shows the confusion matrix, and Table 4 shows the recall, precision and F-score of each age class. It can be seen that the first two classes obtained very good classification results, but the accuracy gradually reduced for the more advanced ages. This suggests that as the tissue structure deteriorates, it becomes harder to differentiate images of similar chronological age.
Table 4: Recall, precision and F-score of age classification.

<table>
<thead>
<tr>
<th></th>
<th>Day 0</th>
<th>Day 2</th>
<th>Day 4</th>
<th>Day 6</th>
<th>Day 8</th>
<th>Day 10</th>
<th>Day 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall (%)</td>
<td>98.1</td>
<td>90.8</td>
<td>62.9</td>
<td>61.4</td>
<td>67.7</td>
<td>56.5</td>
<td>38.9</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>98.1</td>
<td>88.0</td>
<td>64.9</td>
<td>58.4</td>
<td>71.7</td>
<td>46.7</td>
<td>51.2</td>
</tr>
<tr>
<td>F-score (%)</td>
<td>98.1</td>
<td>89.4</td>
<td>63.9</td>
<td>59.8</td>
<td>69.7</td>
<td>51.1</td>
<td>44.2</td>
</tr>
</tbody>
</table>

Figure 8: The distribution of derived morphological ages corresponding to each chronological age, computed with the default parameters $\beta = 10$ and $\delta = 1$.

ages. Especially for the last class, 50% of the images were misclassified as Day 10. This was because the visual characteristics of these two ages were very similar and our method tended to classify these images to the earlier age. It is however encouraging to see that misclassification mainly occurred between neighboring classes rather than the more distant classes. This indicates that our method is effective in identifying relatively clear visual distinctions for images of quite different ages. Also, the misclassification between neighboring classes could be due to the actual visual similarities between images of similar morphological ages but different chronological ages.
4.2. Morphological Age Quantification

The distributions of derived morphological ages are shown in Figure 8. It can be seen that for each chronological age, the majority of images of that chronological age exhibit similar morphological ages with some offsets from their chronological age. For example, the morphological ages obtained for the majority of class Day 2 are within the range of 1.7 and 2.6 days. These correspond to offsets of -0.3 to 0.6 from the chronological age of 2 days. The degree of offsets was controlled by the scaling parameter $\beta$ in Eq. (11). In our study, we used $\beta = 10$ to have large offsets, so that within the same chronological age, there would be clear variation in morphological ages to represent the difference in morphological characteristics. A smaller $\beta$ (e.g. 2) would lead to smaller offsets as shown in Figure 9. In addition, if five classification scores were used to compute the morphological age, i.e. $\delta = 2$ in Eq. (12), the derived morphological ages would become less centralized at the classified chronological ages and hence the distribution of morphological ages for a chronological age would become more continuous. For example, the distribution of morphological ages of class Day 8 in Figure 10 has a more continuous span and the lower and upper ends of the distribution shifted towards the median at 7.5 days, compared to Figure 8.

Figures 8, 9 and 10 also show that there are overlaps in the morphological ages corresponding to each chronological age, with scaling parameter $\beta = 2$ and number of neighboring classes $\delta = 1$.
ages for images of different chronological ages. The overlaps are especially large for the last two age classes, and they correspond with the misclassification rates between the two classes as shown in Table 3. On the other hand, the overlaps in morphological ages could be meaningful since we expect that due to the differences in individual aging, images of the same chronological ages should exhibit varying morphological ages and images of different chronological ages can exhibit similar morphological ages. Such heterogeneity in aging effects between individuals has been evident in prior research (Herndon et al., 2002). However, at this stage, there is no ground truth available for the morphological age, since manual quantification of morphological age has only been experimental and subjective, and is especially difficult for the terminal bulb aging database due to the complex visual characteristics. Nevertheless, the literature in aging studies reports that there is statistically significant correlation between the average level of tissue deterioration and chronological age (Garigan et al., 2002; Helfrich et al., 2007). This implies that there should be strong correlation between the morphological and chronological ages of tissue. Therefore, we suggest that the morphological age should be close to its chronological age.

To this end, we propose to use the root-mean-square error (RMSE) between the derived morphological ages and the annotated chronological ages.
Table 5: RMSE between the derived morphological ages and the annotated chronological ages with various parameter settings in our method or SVM.

<table>
<thead>
<tr>
<th></th>
<th>$\beta = 2$</th>
<th>$\beta = 2$</th>
<th>$\beta = 5$</th>
<th>$\beta = 5$</th>
<th>$\beta = 10$</th>
<th>$\beta = 10$, $\delta = 1$</th>
<th>$\delta = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>1.305</td>
<td>1.366</td>
<td>1.332</td>
<td>1.318</td>
<td>1.357</td>
<td><strong>1.276</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: RMSE between the derived morphological ages and the annotated chronological ages comparing different methods.

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>Lasso</th>
<th>Elastic net</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>1.618</td>
<td>3.311</td>
<td>3.482</td>
</tr>
</tbody>
</table>

as the evaluation metric. Table 5 summarizes the RMSE values when different settings of $\beta$ and $\delta$ are used. Using our method, the lowest RMSE was obtained with $\beta = 10$ and $\delta = 2$ (Figure 10). A further analysis shows that with $\beta = 10$, lower RMSEs were obtained with $\delta = 1$ than $\delta = 2$ for the first two classes, but higher RMSEs were obtained for the other classes; hence the overall RMSE was lower with $\delta = 2$. We also found that when $\delta = 3$ was used to compute the morphological age using the classification scores of all seven classes, the resultant morphological ages deviated much from the classified chronological ages and hence $\delta = 3$ was not used.

We also tested with SVM as the classifier, and computed the morphological ages by combining the probability estimates from SVM and Eq. (12) (with the best setting $\delta = 2$). From Tables 5 and 6, it can be seen that our method obtained lower RMSE than SVM, regardless of the parameter settings. In addition, least-squares regression methods with Lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005) regularization were evaluated. We consider that age quantification can be intuitively formulated as a least-squares regression problem due to the continuous nature of age. The regression model predicts the morphological age based on the image feature descriptor and the coefficients learned using Lasso or elastic net. The parameters were tuned manually to achieve optimal performance. Table 6 shows that the RMSEs obtained using these regression methods were much higher than our approach and SVM. Therefore, by first optimizing the classification
Table 7: Pearson correlation between the morphology quantification and the annotated chronological ages comparing different methods.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>0.924</td>
<td>0.54</td>
<td>0.69</td>
</tr>
</tbody>
</table>

of chronological ages with DDLC, our method could predict a morphological age that is more closely correlated with the chronological age, compared to the popular least-squares regression techniques.

4.3. Interpretation of Tissue Morphology

There are two notable approaches in the literature that support automated morphological age quantification (Johnston et al., 2008; Shamir et al., 2009), which also use the same terminal bulb aging images as our study. The overall method of the first approach (Johnston et al., 2008) is relatively similar to ours. It extracts a high-dimensional feature descriptor and uses Fisher discrimination and distance-based computation to derive a morphological age score for each image. The second approach (Shamir et al., 2009) investigates the effects of texture features (entropy and directionality) in representing the age morphology. Both approaches used the Pearson correlation between the quantification results (age score or texture value) and the tissue chronological ages as a quantitative evaluation metric; and suggested that a high Pearson correlation would indicate that the method was effective in capturing the morphological characteristics, based on the hypothesis that the morphological age should be closely correlated with the chronological age.

Table 7 shows the Pearson correlation of our method and those reported by the existing approaches. All methods had p-values much smaller than 0.05, indicating the correlations between the quantification results and chronological ages were statistically significant. Our method achieved a much higher Pearson correlation than the compared approaches. This implies that our method could identify subtle morphological differences between different age classes and better distinguish the various stages in tissue aging. This observation can be further demonstrated by Figure 11. Following the approach (Johnston et al., 2008), one tailed t-test was performed to assess if the derived morphological ages were different between images of different chronological ages with statistical significance. Using our method, a p-value of 0.022 was
obtained when comparing the morphological ages from Day 10 and Day 12, and p-value \( < 10^{-19} \) was obtained for all the other cases. On the other hand, the p-values reported (Johnston et al., 2008) were greater than 0.05 for six pairs of chronological ages, as shown in Figure 11.

Overall, the existing studies (Johnston et al., 2008; Shamir et al., 2009) suggest that there are three states of morphological ages based on the quantification results. Specifically, tissues of Day 0 and Day 2 belong to the first state, tissues of Day 4, Day 6, and Day 8 are the second state, and Day 10 and Day 12 as the third state. There are relatively clear transitions between different states and the quantification values are similar for images of the same state. When using our method, while there are also evident overlaps in morphological ages for tissues of Day 4, Day 6, and Day 8, and for tissue of Day 10 and Day 12 (Figure 8), such overlaps are relatively small and the differences are statistically significant (Figure 11). Therefore, our finding is that tissues of different chronological ages exhibit distinguishable morphological characteristics, and there is a continuous trend of morphological changes as the tissue ages as opposed to three states.

For a more clinical assessment of the quantification results in representing the morphological changes, one way is to evaluate if normal and pathological tissues exhibit different morphological ages even if their chronological ages
are the same. The existing study (Johnston et al., 2008) has performed this analysis and concluded the effectiveness of their method in capturing such morphological changes. Although we do not have such data for analysis, we hypothesize that our method would be effective based on the following reasoning. Consider that our method encodes the similarity of an image with the typical characteristics of each chronological age. Due to the pathology, the tissue would appear visually similar to the more advanced age class. Our method would thus derive a morphological age that is more different from its chronological age, and such a morphological age would effectively represent the morphological changes from the normal tissue.

In addition, we would like to mention that different from the existing approach (Shamir et al., 2009) that evaluates texture entropy and directionality, our method does not produce an intuitive description of the morphological features. We can consider that our method generates a seven-dimensional DDLC coding, with each dimension defined as the classification score in Eq. (11). This DDLC coding is then an abstract feature representation of the image. Although this abstract nature causes difficulties to comprehend the feature and interpret the underlying morphology of tissue aging, our method provides an alternative to model-based analysis (Shamir et al., 2009) and enables aging research based on a quantitative morphological age. For example, our method can be used in place of the manual grading of morphological ages and assess the correlation between various biological or environmental factors and tissue aging.

Figure 12: Pearson correlation between the DDLC coding and the chronological age.
To evaluate the descriptiveness of our feature (the DDLC coding), we measured the Pearson correlation between the feature and the chronological age, following the evaluation protocol (Shamir et al., 2009). As shown in Figure 12, the Pearson correlation was negative for the first three dimensions of the DDLC coding, and had an increasing trend among the various dimensions. These results are as expected since there should be a negative correlation between the tissue age and the classification scores of the younger age classes, and a strong positive correlation for the advanced age classes. The misclassification between last two classes caused a drop in Pearson correlation for the last dimension of the DDLC coding. Note that all correlations were statistically significant with p-value $< e^{-19}$.

5. Conclusions and Future Work

We present an automated method for quantifying the morphological ages of tissues. Our method consists of three steps. First, image features are extracted at two scales using multiple types of descriptors. Second, the chronological age of tissue is classified using our dual discriminative local coding (DDLC) model, which is designed based on the locality-constrained local coding method with additional discriminative distance learning and dual-level local coding to enhance the discriminative capability. Finally, based on the classification outputs, the morphological age of tissue is computed. Our study is conducted on the publicly available terminal bulb aging database, which comprises 970 DIC microscopy images of seven different chronological ages. Our evaluation shows that our feature design and DDLC classifier achieve significant improvement over the existing approaches on chronological age classification. We also show that our exploratory design generates morphological ages that are closely correlated with the chronological ages, and can be a feasible alternative to manual age estimation.

For future work, we will study methods to improve the classification performance especially for Day 10 and Day 12. We suggest that the current limitation mainly lies in the feature description, and hence more comprehensive features based on manual design or automated learning will be investigated. We will also make our software available to researchers on tissue aging, so that the effect of our method in supporting aging research can be systematically evaluated.
6. Acknowledgement

This work was supported in part by Australian Research Council (ARC) grants.

References


