ABSTRACT Tongue diagnosis plays a key role in TCM (Traditional Chinese Medicine) diagnosis. Tongue image segmentation lays a solid foundation for quantitative tongue analysis and diagnosis. However, the segmentation of tongue body is challenging due to the factors such as large personal variation of tongue body on color, texture, and shape, as well as weak edges caused by similar color between tongue body and neighboring tissues, especially the lip. Existing segmentation methods usually use only single color component and simple prior knowledge, thus leading to inaccuracy and instability. To alleviate these issues, a patch-driven segmentation method with sparse representation is proposed in this paper. Specifically, each patch in the testing image is sparsely represented by patches in the spatially varying dictionary, which is constructed by the local patches of training images. The derived sparse coefficients are then employed to estimate the tongue probability. Finally, the hard segmentation is obtained by applying the maximum a posteriori (MAP) rule on the tongue probability map and further polished with morphological operations.

INDEX TERMS Chinese medicine, image segmentation, medical image processing, sparse representation, tongue diagnosis.

I. INTRODUCTION
Tongue diagnosis [1], [2] aims to draw physiological and pathological cues according to tongue body’s features including color, texture, shape, coating and so on. Tongue diagnosis has the virtues of effectiveness, painlessness, no side-effects, simplicity and immediacy, which make it be one of the most important and widely used diagnostic methods in the TCM. After practices over 3000 years, and TCM practitioners have accumulated very rich clinical experiences such as eight classic tongue diagnosis principles [2] revealing strong relationships between tongue body’s sub-regions and human organs’ health statuses. For examples, tongue coating is usually strongly related with the health status of stomach. In addition, tongue’s appearance can be used to monitor a patient’s health status. In detail, the color and texture of tongue coating usually reflect many diseases and human health statuses, such as inflammation, infection, and stress [3]. The researchers of TCM tongue diagnosis have also explored value of tongue diagnosis in diagnosing diseases, such as cancer [4], gastritis, and precancerous lesions [5].

In the modern Western medicine, some literatures [6], [7] named tongues with discolored regions or cracks as “geographic tongue”, and validated the value of the quantitative tongue diagnosis for making clinical decision [7]. For example, tongue coating exhibits strong relationship with viable salivary bacteria [8], which indicates a risk of aspiration pneumonia in edentate patients. Amyloidosis of the tongue may be used to diagnose...
plasmacytoma [9]. The research [10] explored connection between tongue diagnosis and tongue-coating microbiome, systematically investigating the biological bases of different tongue-coating appearances at the molecular level. The above examples exhibit the potential of tongue diagnosis for inferring systemic disorders.

However, the accuracy of conventional tongue diagnosis highly depends on the practitioners’ experiences, and the diagnosis has subjective and time-consuming limitations. To solve these issues, some image processing and pattern recognition techniques were used to achieve automatic tongue diagnosis [11]. This type of automatic diagnosis usually first uses image segmentation technique to extract tongue body in a tongue image, then uses feature extraction technique to calculate tongue body features, and finally utilizes a classifier to achieve final diagnosis. Hence, tongue image segmentation is a crucial step of achieving automatic tongue diagnosis. Some works [12]–[16] have been done to resolve this problem. However, tongue image segmentation is currently still challenging due to large personal variation of tongue body appearances, and weak edges mainly caused by similar color between tongue body and lip.

In recent years, patch-based methods draw more and more attentions in the fields of computer vision and image processing, such as label fusion and segmentation [17]–[20], texture synthesis [21], in-painting [22], image denoising [23], [24], and super-resolution reconstruction [25]. For example, Coupé et al. [17] presented a patch-based label fusion method termed as the nonlocal-means method for segmentation of hippocampus and ventricle. This patch-based method is widely used due to its simplicity and good segmentation performance. Recently, patch-based sparse representation has also attracted increasing interest in the field of image processing. These patch-based methods assume that each patch (i.e., local window) in an image can be sparsely represented by a linear combination of image patches in an over-complete dictionary [26]–[32]. Motivated by the nonlocal-means method [17] and the sparse representation techniques, we propose a novel patch-based sparse representation method for tongue image segmentation in this paper. For the patch-based segmentation, the differences between the nonlocal-means method [17] and our proposed method include two aspects: 1) the nonlocal-means method [17] presented a similarity measure in grayscale space to pre-select the training patches. The proposed method extends the similarity measure in [17] from grayscale space to RGB color space. 2) The nonlocal-means method took the weighted average of training pixels’ labels as the probability of the testing pixel belonging to the object (tongue body), where the weight of a training pixel is defined as the similarity between its corresponding training patch and the testing patch. Different from the nonlocal-means method, the proposed method uses sparse coefficients obtained by sparse representation to calculate the probability of the testing pixel belonging to the object. In summary, for the nonlocal-means method, all the remaining training patches after patch pre-selection contribute to the estimation of object’s probability, regardless of their similarity to the testing patch. However, for the proposed method, only the training patches with non-zero sparse coefficients in the dictionary contribute to the estimation of object’s probability. Specifically, the proposed method first uses the pre-selected local patches of training images (through similarity-based patch pre-selection) to construct a dictionary. Then, for each pixel in a testing image, its corresponding patch can be sparsely represented by the patches in the dictionary, and thus this patch will have similar segmentation label as the selected patches in the dictionary according to their respective sparse coefficients. By visiting each pixel in the testing image and computing the respective probability of corresponding image patch, a tongue probability map can be generated for the entire testing image. Lastly, the final hard segmentation can be obtained by applying the maximum a posteriori (MAP) rule on the tongue probability map and further polishing with morphological operations. Experimental results on 290 subjects and additional 10 unseen subjects show that the proposed method can dramatically improve both segmentation accuracy and robustness.

The rest of this paper is organized as follows. Related work is introduced in Section II. Theory and implementation of the proposed method is described in Section III. Experimental results are reported in Section IV. Conclusions are drawn in Section V.

II. RELATED WORK

Due to the above challenges of tongue image segmentation, conventional simple low-level image processing techniques such as edge detection fail to obtain satisfactory segmentation result. To improve accuracy of tongue image segmentation, some hybrid tongue image segmentation works were developed, where ACM (Active Contour Model) based methods are popular. ACM also called snake [33] evolves a given initial curve near to true object contour under controls of internal and external forces.

ACM-based methods mainly study on initial contour determination and curve evolution. For example, Pang et al. [13] presented a bi-elliptical deformable contour extraction method, which is termed as BEDC. The method [13] first defined a specific bi-elliptical deformable template (BEDT) as a rough description of tongue body, then obtained the initial tongue contour by minimizing the BEDT energy function, and finally used a modified active contour model of replacing the conventional internal force with the template force to evolve the initial contour and obtain the final segmentation result. BEDC [13] has two limitations. One limitation is that BEDC used only the red component of a tongue image. As pointed out in [15], color is usually the most important feature to distinguish tongue body from image background. But single color component cannot adequately use color information, thus easily generating the confusion between tongue body and its neighboring tissues. This further adds the difficulty of tongue image segmentation on the weak edges. The other limitation is that BEDC [13] simply used two
FIGURE 1. Flowchart of the proposed method.

semi-ellipses to model the shape of tongue body. However, the rough modeling and the large variation of tongue shape may make the initial tongue contour contain the undesirable strong edges from neighboring tissues, thus converging far away from the true tongue contour.

On the other hand, Zhang et al. [14] used ACM and polar edge detection to achieve tongue image segmentation. Specifically, it first detects tongue body boundaries via polar edge detection, and removes fake tongue body boundaries via an edge mask. Then, it performs a local adaptive bi-thresholding and morphological filtering for converting the edge filtering result to a bi-level polar edge image. Finally, it obtains the initial tongue body contour in a heuristic way, and refines the contour via ACM. But, the method failed to resolve two main challenges in tongue image segmentation, i.e., 1) how to extract weak tongue body contour caused by color similarity between the tongue body and its surroundings, and 2) how to distinguish edges belonging to tongue body contour from other edges caused by strong textures or tongue body coating. To remove non-tongue body edges, the method [14] used Sobel operator, Gaussian filter, image thresholding, and morphological operation. But the scheme is invalid for those long non-tongue body edges. In addition, Gaussian filter can often weaken the true tongue contour, thus making the first challenge more difficult to solve.

Ning et al. [15] used gradient vector flow (GVF), region merging (RM), and ACM to extract tongue body. The method is briefly termed as GVF-RM. Specifically, this method first improved GVF for suppressing both noise and trivial image structures, then obtained many small image regions via watershed segmentation, and finally performed region merging to construct the initial tongue body contour for subsequent ACM-based contour refinement. GVF-RM [15] has three limitations. First, it used only red component during obtaining the initial contour of tongue body. Second, when performing initial tongue body contour determination based on region merging, it obtained object (tongue body) region marker and background region marker under an assumption that tongue body and background are at image center and image border, respectively. However, when tongue body is very close to image border, the wrong background marker may cause the above assumption be invalid, easily generating bad segmentation result. Third, when it improved GVF to suppress some noise and trivial image structures, it also weakened true tongue body contour.

Furthermore, the above ACM-based methods are usually sensitive to the initial object contour. Once the initial object contour includes fake strong edges from neighboring tissues such as neck, cheek, face, wrinkle and lip, the initial contour is difficultly converged to true boundary of tongue body.

III. METHOD

To address limitations of existing tongue image segmentation methods, inspired by patch-based sparse representation, this paper proposes a novel patch-driven method, and Fig. 1 shows flowchart of the proposed method. The contributions of our proposed method are as follows:

1) We innovatively introduce a patch-based sparse representation into tongue image segmentation. The proposed method adequately uses prior information from the training tongue images with manual segmentations, thus dramatically improving both segmentation accuracy and robustness.

2) We develop a novel color information based criterion to evaluate the similarity between a testing patch and a training patch, and then use it to pre-select training patches for reducing the size of the dictionary in sparse representation and also saving the computational time.

A. DICTIONARY CONSTRUCTION

Our proposed method converts the problem of tongue image segmentation as a pixel classification task. Specifically, it uses a patch-based sparse representation to calculate the probability of each testing pixel belonging to the object (tongue body), and then obtains the final segmentation result by subsequent morphological operations. During conducting the patch-based sparse representation, the size of the overcomplete dictionary, constructed by training image patches, could seriously affect the computational time. Also, including those confusing patches into the dictionary could affect the sparse representation result and eventually the segmentation result. To address this issue, we design a new similarity measure to perform a pre-selection of training patches for each testing patch. The designed new measure extends the conventional similarity measure [17] (evolved from the well-known structural similarity measure (SSIM) [34]) from grayscale space into color space. The definition of our new similarity
measure is as follows:

\[
SIM(S_x, P_y) = \frac{1}{3} \sum_{c=1}^{3} \left( \frac{2\mu_{S_x,c}\mu_{P_y,c}}{\mu_{S_x,c}^2 + \mu_{P_y,c}^2} + \frac{2\sigma_{S_x,c}\sigma_{P_y,c}}{\sigma_{S_x,c}^2 + \sigma_{P_y,c}^2} \right),
\]

(1)

where \(S_x\) and \(P_y\) denote the testing patch and the training patch centered at the locations of pixel \(x\) and pixel \(y\), respectively. \(c\) indicates the index of the RGB color components, where each component has 8-bit depth with the intensity range of \([0, 255]\). \(\mu\) and \(\sigma\) represent the mean and standard deviation of the patch from the corresponding color component. High \(SIM\) value means the high similarity between \(S_x\) and \(P_y\).

The above measure evaluates the similarity between the testing patch \(S_x\) and the training patch \(P_y\), which can be also used to pre-select training patches. The criterion of our training patch pre-selection is formulated as:

\[
B(S_x, P_y) = \begin{cases} 
1, & \text{if } SIM(S_x, P_y) > th \\
0, & \text{otherwise}
\end{cases}
\]

(2)

where 1 means that the training patch \(P_y\) has high similarity with the testing patch \(S_x\), thus \(P_y\) should be included into the dictionary for sparse representation. Conversely, 0 indicates low similarity between patches \(P_y\), and \(S_x\), thus \(P_y\) should be excluded. The parameter \(th\) was set to 0.95 for all the experiments according to the literature [17]. Note that, to avoid repeated computations and save computational time, both mean and variance of each patch can be pre-calculated as maps of local means and local variances.

To construct a patch dictionary for each pixel \(x\) in a testing image \(I\), the color patch centered at the pixel \(x\) (taken from \(w \times w\) neighborhood and the RGB color components) is first columnized to be a \(3w^2\) dimensional vector, and this vector is then normalized to have the unit L2 norm [30], denoted as \(S_x\). Furthermore, the patch dictionary for the pixel \(x\) can be adaptively built from all training images with manual segmentation maps as follows. First, letting \(\Omega(x)\) be the \(w_x \times w_y\) neighborhood of pixel \(x\) in the \(i\)-th training image, we can obtain the patch for each pixel \(y \in \Omega(x)\) from the training image, and pre-select all patches from \(w_x \times w_y\) neighborhoods of all training images according to the criterion defined in Eq. (2). Then, each selected training patch is columnized and sequentially normalized as a \(3w^2\) dimensional column vector \(P_y\) with the unit L2 norm. Finally, the normalized column vectors of all selected training patches are stacked to be a \(3w^2 \times N\) matrix as the dictionary, \(D_x\), for sparse representation, where \(N\) denotes the total number of selected training patches. For better clarity, Fig. 2 shows the flowchart of dictionary construction, where \(x\) indicates the testing pixel, yellow squares in training images are the searching neighborhoods centered at the location of \(x\), and squares with other colors in training images are the training patches.

B. SPARSE REPRESENTATION

To represent the testing patch \(S_x\) by the dictionary \(D_x\), the sparse coefficient vector \(\alpha\) can be estimated by minimizing the following non-negative Elastic-Net problem [35],

\[
\min_{\alpha \geq 0} \frac{1}{2} \|S_x - D_x \alpha\|_2^2 + \lambda_1 \|\alpha\|_1 + \frac{\lambda_2}{2} \|\alpha\|_2^2.
\]

(3)

In Eq. (3), the first term is the data fitting term, the second term is the L1 regularization term enforcing sparsity of \(\alpha\), and the last term is the L2 smoothness term enforcing similarity of coefficients for similar image patches. Eq. (3) is a convex combination of L1 lasso [36] and L2 ridge penalty, which encourages a grouping effect while keeping a similar sparsity of representation [35]. Our implementation used the LARS algorithm [37] in the SPAMS toolbox (http://spams-devel.gforge.inria.fr) to solve the Elastic-Net problem. Each element of \(\alpha\), i.e., \(\alpha_i\), reflects similarity between the testing patch \(S_x\) and the training patch \(P_y\) in \(D_x\) [38].

Based on the assumption that similar patches should share similar segmentation labels, sparse coefficients in \(\alpha\) are used to estimate probability of the pixel \(x\) belonging to the object (tongue body):

\[
p(x) = \frac{1}{Z(x)} \sum_{i=1}^{N} \alpha_i L(P_y^i).
\]

(4)
where
\[
Z(x) = \sum_{i=1}^{N} \alpha_i, \tag{5}
\]
is a normalization constant, and \(L(P_i^y)\) is the manual segmentation label of the central pixel of the \(i\)-th training patch \(P_i^y\), where 1 and 0 are used for the object label and the background label, respectively.

After calculating each testing pixel’s probability, a probability map can be obtained. Fig. 3 shows the probability maps obtained for eight representative tongue images. These eight images represent large variation of tongue body on shape, size, color, texture, and coating. For example, Figs. 3(a)-(b) show the representative examples with large tongue shape variation. From Fig. 3, it can be seen that the proposed method generates accurate probability maps. To convert a probability map into a hard segmentation, the label of the pixel \(x\) can be determined by using the maximum a posteriori (MAP) rule. As shown in Fig. 4, the hard segmentation (Fig. 4(c)) is converted from the estimated probability map (Fig. 4(b)). However, there may be some artifacts, such as isolated faked objects and holes as indicated by the blue and red arrows in Fig. 4(c).

These artifacts can be easily corrected by morphological operations. For example, those isolated faked objects can be removed by the area opening operation, while holes can be filled by the morphological closing operation. Final segmentation result is shown in Fig. 4(d) after performing these morphological operations.
IV. EXPERIMENTAL RESULTS

To validate the effectiveness of the proposed method on tongue image segmentation, we conducted extensive experiments on a color tongue image dataset including 290 subjects by 10-fold cross-validation, as well as additional 10 unseen subjects. Segmentation results on 8 representative tongue images were used to perform qualitative comparison between the proposed method and state-of-the-art methods, i.e., GVF-RM [15] and a popular patch-based label fusion method called nonlocal-means [17]. Then, segmentation results on the entire tongue image dataset were used to perform quantitative comparison evaluated by 4 common classification measures [39]–[41]. And finally, segmentation results on additional 10 unseen images were qualitatively and quantitatively compared.

Based on the highest average kappa index (KI) value [41] on the entire image dataset, GVF-RM determined the optimal iteration number for GVF-based image diffusion, and then taken corresponding results as final segmentation results for quantitative and qualitative comparison with other methods. Other parameters of GVF-RM follow the original literature [15]. The non-local-means label fusion method [17] shares the same patch pre-selection and experimental setting of 10-fold cross-validation as the proposed method.

A. TONGUE SUBJECTS AND DATA ACQUISITION

Tongue images were acquired by a high quality and consistent tongue imaging system [42] with the flowchart given in Fig. 5, which was designed by the Hong Kong team of Prof. David Zhang. Each image has the size of 160 × 120 and the resolution of 1 × 1 mm² in our experiments. The manual segmentations were performed by the collaborators of the Hong Kong team, i.e., Dr. Naimin Li and his team from Harbin Binghua Hospital.

B. EVALUATION METRICS

To quantitatively evaluate image segmentation accuracy, we used 4 classification measures, i.e., misclassification error (ME) [39], false positive rate (FPR), false negative rate (FNR) [40], and kappa index (KI) [41], where

\[ ME = 1 - \frac{|B_m \cap B_a| + |F_m \cap F_a|}{|B_m| + |F_m|}, \]  
\[ FPR = \frac{|B_m \cap F_a|}{|B_m|}, \]  
\[ FNR = \frac{|F_m \cap B_a|}{|F_m|}, \]  
\[ KI = 2 \frac{|F_m \cap F_a|}{|F_m| + |F_a|}. \]

\( B_m \) and \( F_m \) indicate background and foreground in manual segmentation result (ground truth), \( B_a \) and \( F_a \) indicate background and foreground in an automatic algorithm’s segmentation result, and |·| is cardinality of a set. The values of ME, FPR, FNR, and KI are between 0 and 1, lower values of the front three measures, and higher KI value indicate better segmentation accuracy.

C. IMPACTS OF THE PARAMETERS

The proposed method has four important parameters, i.e., \( w \), \( w_l \), \( \lambda_1 \) and \( \lambda_2 \). Parameters \( w \) and \( w_l \) denote the patch size and the size of neighborhood, respectively. Parameters \( \lambda_1 \) and \( \lambda_2 \) are the weights of L1 regularization term and L2 smoothness term in sparse representation. Four parameters were adaptively determined via 10-fold cross-validation in our experiments. To clearly discuss the impacts of four parameters on the segmentation accuracy of the proposed method, we fixed two parameters and varied other two parameters. Specifically, we first empirically fixed \( \lambda_1 \) and \( \lambda_2 \) in Eq. (3) to 0.1 and 0.01, and evaluated the impacts of \( w \) and \( w_l \) on the segmentation accuracy. Then, we fixed \( w \) and \( w_l \) to 7 and 19, respectively,
according to the evaluation results of \( w \) and \( w_l \), and tested the impacts of \( \lambda_1 \) and \( \lambda_2 \) on the segmentation accuracy. To observe the impacts of the parameters on the segmentation accuracy, we specifically recorded the average kappa index values of all the images from all the validation sets in 10-fold cross-validations under different parameter combinations.

First, we studied the impacts of patch size (\( w \)) and search neighborhood size (\( w_l \)) on the segmentation accuracy, where \( w \) is selected from \{3, 7, 11, 15, 19\}, and \( w_l \) is selected from \{7, 11, 15, 19\}. The kappa index results are provided in Fig. 6(a) and Fig. 6(b), respectively. Fig. 6(a) shows that the best average kappa index values were obtained with a patch size of \( 7 \times 7 \) for each fixed \( w_l \), where the best values are 0.961, 0.964, 0.967, and 0.969 for \( w_l = 7, 11, 15 \), and 19, respectively. The optimal patch size seems to reflect the complexity of the tongue structure. That is, the patch size needs to be large enough for description of tongue structure, but should not be too large to miss the localization. Fig. 6(b) shows that the best average kappa index values were obtained with a search neighborhood of \( 19 \times 19 \) for each fixed \( w \), where the best values are 0.966, 0.969, 0.968, 0.966, and 0.964 for \( w = 3, 7, 11, 15, \) and 19, respectively. As expected, the performance is improved with the increase of the size of \( w_l \).

We also studied the impacts of \( \lambda_1 \) and \( \lambda_2 \) on the segmentation accuracy, where \( \lambda_1 \) and \( \lambda_2 \) are selected from \{0, 0.01, 0.1, 0.2, 0.5\}. The kappa index results are provided in Fig. 6(c) and Fig. 6(d), respectively. Fig. 6(c) shows the best average kappa index values obtained with \( \lambda_1 = 0.1 \) for each fixed \( \lambda_2 \), where the best values are 0.968, 0.969, 0.965, 0.964, and 0.962 for \( \lambda_2 = 0, 0.01, 0.1, 0.2, \) and 0.5, respectively. Fig. 6(d) shows the best average kappa index values obtained with \( \lambda_2 = 0.01 \) for each fixed \( \lambda_1 \), where the best values are 0.968, 0.968, 0.969, 0.968, and 0.966 for \( \lambda_1 = 0, 0.01, 0.1, 0.2, \) and 0.5, respectively. From Figs. 6(c) and (d), it can also be observed that, if there is no sparse constraint (\( \lambda_1 = 0 \)) and smooth constraint (\( \lambda_2 = 0 \)), which means that all reference patches in the dictionary could contribute to the estimation of object’s probability, regardless of their similarity to the testing patch, the kappa index value (0.939) is the lowest among all combinations of \( \lambda_1 \) and \( \lambda_2 \). The accuracies under other combinations of \( \lambda_1 \) and \( \lambda_2 \) are significantly improved, with the best accuracy obtained with \( \lambda_1 = 0.1 \) and \( \lambda_2 = 0.01 \). This demonstrates the importance of using both L1 regularization term and L2 smoothness term in Eq. (3). On the other hand, there is no significant difference among different combinations of \( \lambda_1 \) and \( \lambda_2 \), except \( \lambda_1 = 0 \).
and \( \lambda_2 = 0 \), indicating the robustness of our proposed method to the values of \( \lambda_1 \) and \( \lambda_2 \). However, the above ranges for four parameters were empirically chosen in our experiments, which could lead to local minimum results.

### D. THE IMPACT OF COLOR SPACE

This section discusses the impact of color space on the segmentation accuracy of the proposed method. First, we studied the impacts of different color spaces on the segmentation accuracy, where color spaces include RGB and CIELAB. When applying the proposed method to LAB color space, we first converted each image from RGB to LAB by using the inherent function “rgb2lab” in MATLAB 2015a. Then, we performed experiments on the entire image dataset in LAB color space, including 290 subjects, by 10-fold cross-validation. Our experiments in LAB color space share the same patch pre-selection and experimental setting of 10-fold cross-validation as the experiments performed in RGB color space. The only difference is to use LAB color space, instead of RGB color space. The results by LAB and RGB color spaces are shown in Fig. 7(a). The mean and standard deviation of kappa index (KI) values in LAB and RGB spaces are 0.960 ± 0.036 and 0.972 ± 0.023, respectively. These quantitative results show that the proposed method obtains significantly higher segmentation accuracy (p-value = 7.557 \times 10^{-5}) and also the stronger stability in RGB color space, in comparison with LAB color space.

Fig. 7(b) shows the comparison of average KI values obtained by applying the proposed method to the entire image dataset in both grayscale space and RGB color space by 10-fold cross-validation. The mean and standard deviation of KI values obtained by the proposed method in the grayscale space and RGB color space are 0.934 ± 0.055 and 0.972 ± 0.023, respectively. These quantitative results show that the usage of color information indeed improves the segmentation accuracy of the proposed method.

### E. QUALITATIVE COMPARISON

To qualitatively compare segmentation effects of three methods, Fig. 8 shows their segmentation results on the eight representative tongue images with large variation of tongue body on shape, size, color, texture, and coating. GVF-RM obtains satisfactory segmentation result only on Fig. 8(e), but generates misclassification on all other images. Specifically, GVF-RM suffers from heavy misclassification on Fig. 8(a). Also, GVF-RM generates under-segmentation (see examples in Figs. 8(b)-(d) and (f)-(h), where the true tongue body contours are pointed out by the red arrows). On the other hand, GVF-RM generates over-segmentation (see examples in Figs. 8(a)-(b) and (h), where the true tongue body contours are pointed out by the green arrows). Similarly, the nonlocal-means method suffers from misclassification on most of images. For example, the nonlocal-means method either suffers from under-segmentation as indicated in Figs. 8(b)-(c) and (e)-(f), or generates over-segmentation as indicated in Figs. 8(b), (d) and (g), where the true tongue body contours are also pointed out by the red and green arrows. Compared with the above two methods, the proposed method obtains more accurate segmentation results on all the eight images, and our contours are very close to the true tongue body contours. Experimental results validate the robustness of the proposed method.

### F. QUANTITATIVE COMPARISON

To quantitatively compare segmentation accuracy of GVF-RM, nonlocal-means method, and the proposed method, we evaluated them on the entire dataset using four measures including ME, FPR, FNR and KI. Fig. 9 shows box-whisker plots of the four measures. Furthermore, the mean and standard deviation of ME values obtained by three methods are 0.091 ± 0.091, 0.044 ± 0.035 and 0.012 ± 0.010, respectively. The mean and standard deviation of FPR values obtained by three methods are 0.071 ± 0.105, 0.029 ± 0.027 and 0.006 ± 0.008, respectively. The mean and standard deviation of FNR values obtained by three methods...
are 0.153 ± 0.165, 0.086 ± 0.093 and 0.031 ± 0.030, respectively. These quantitative results show that the proposed method has the lowest misclassification error and strongest stability. For kappa index (KI), three methods achieve segmentation accuracies of 0.806 ± 0.162, 0.899 ± 0.078 and 0.972 ± 0.023, respectively, which again demonstrates superiority of the proposed method.

G. RESULTS ON 10 UNSEEN SUBJECTS
Besides the 10-fold cross-validation experiments on the image dataset including 290 subjects, we further validated our proposed method on additional 10 unseen subjects, with their manual segmentations provided by Dr. Naimin Li from Harbin Binghua Hospital. Fig. 10 shows segmentation results on three of these 10 unseen images. As can be observed that
the proposed method obtained higher segmentation accuracy than GVF-RM [15] and the nonlocal-means method [17], by referring to the manual segmentation results as ground truths. The average kappa index values on 10 subjects by different methods are shown in Fig. 11, which again demonstrates the advantage of the proposed method.

**H. TIME COMPLEXITY**

The total complexity of the proposed method is about $O(n_1(3w^2 \cdot w_l^2 + N_1 \cdot w_l^2 + T_s + m_1))$, where $n_1$ is the pixel number of a testing image, $w \times w$ is the patch size, $w_l \times w_l$ is the search neighborhood size, $N_1$ is the number of training images, $T_s$ is the time complexity of a testing pixel’s local patch sparsely represented by an over-complete dictionary using sparse representation technique, and $m_1$ is the number of the remaining training patches after patch pre-selection and $m_1 \ll N_1 \cdot w_l^2$.

To compare computational efficiency of three methods, we also specifically recorded their average computational times on the entire tongue image dataset with image size of
160 × 120. The average computational time of GVFRM is around 0.2 minutes, while the average computational times of the nonlocal-means method and the proposed method are around 3 minutes and 5 minutes, respectively, under the parameter combination of \( w = 3 \) and \( w_1 = 7 \) in our Linux server with 1 CPU and 2G memory. Under this parameter combination, the proposed method achieves the accuracy of 0.964 ± 0.025, measured by kappa index (KI). As for the computational time of the proposed method, on average, it takes around 2.8 minutes to construct the over-complete dictionary and then additional 2 minutes to perform sparse representation. The rest of the time is consumed by subsequent morphological operations. The proposed method is slower than other two methods, especially the GVFRM. Fortunately, the running time of the proposed method can be significantly shortened by some parallel computing schemes such as modifying the program structure of our algorithm coded by MATLAB, and taking an image segmentation task into multiple sub-image segmentation tasks by decomposing an image into multiple sub-images. For example, when calculating the probabilities of all testing pixels belonging to the object (tongue body) in a testing image, if we respectively decompose the testing image, each training image and corresponding manual segmentation result into 16 sub-images, the running time of the proposed method segmenting an image is about 20 seconds. Using the above parallel computing scheme, the running time of the nonlocal-means method is about 11 seconds accordingly. However, GVFRM is unsuitable for the parallel computing scheme of decomposing an image segmentation task into multiple sub-image segmentation tasks, as it uses the information of the whole image in its image segmentation process.

V. CONCLUSION

In this paper, we propose a novel patch-driven tongue image segmentation method with sparse representation. The proposed method innovatively introduces sparse representation into tongue image segmentation and utilizes prior information from the tongue training images with manual segmentation. The proposed method also presents a new color-based patch pre-selection criterion to reduce the size of training patches during dictionary construction, thus saving the computational time in sparse representation. Experimental results on a tongue image dataset containing 290 subjects and also additional 10 unseen subjects demonstrate a dramatic performance improvement by the proposed method, compared to the state-of-the-art methods.

REFERENCES

[1] Z. Xie, Practical Traditional Chinese Medicine. Beijing, China: Foreign Language Press, 2000.
[2] B. Kirschbaum, Atlas of Chinese Tongue Diagnosis. Seattle, WA, USA: Eastland, 2000.
[3] C.-H. Hsu, M.-C. Yu, C.-H. Lee, T.-C. Lee, and S.-Y. Yang, “High eosinophil cationic protein level in asthmatic patients with ‘Heat’ zheng,” Amer. J. Chin. Med., vol. 31, no. 2, pp. 277–283, Apr. 2012.
[4] Q. Liu, “Quantitative study on tongue color in primary liver cancer patients by analysis system for comprehensive information of tongue diagnosis,” J. Chin. Integrative Med., vol. 1, no. 3, pp. 180–183, Sep. 2003.
[5] C.-D. Li, Q. F. Lan, and J. J. Zhang, “Study on relationship between tongue picture and cell apoptosis in patients with chronic gastritis,” (in Chinese), J. Chin. Integrative Med., vol. 23, pp. 433–435, Jun. 2003.
[6] Y. Zadik, S. Drucker, and S. Pullmon, “Migratory stomatitis (ectopic geographic tongue) on the floor of the mouth,” J. Amer. Acad. Dermatol., vol. 65, no. 2, pp. 459–460, 2011.
[7] J. K. Anastasi, L. M. Currie, and G. H. Kim, “Understanding diagnostic reasoning in TCM practice: Tongue diagnosis,” Alternative Therapies Health Med., vol. 15, no. 3, pp. 18–28, 2009.
[8] S. Abe, K. Ishihara, M. Adachi, and K. Okuda, “Tongue-coating as risk indicator for aspiration pneumonia in edentate elderly,” Arch. Gerontol. Geriatrics, vol. 47, no. 2, pp. 267–275, Sep. 2008.
[9] S. Hoeft, E. Schilling, S. Philippou, and H. Eufinger, “Amyloidosis of the tongue as a possible diagnostic manifestation of plasmacytoma,” (in German), Mund Kiefer Gesichts chir., vol. 3, no. 1, pp. 46–49, 1999.
[10] B. Jiang, X. Liang, Y. Chen, T. Ma, L. Liu, J. Li, R. Jiang, T. Chen, X. Zhang, and S. Li, “Integrating next-generation sequencing and traditional tongue diagnosis to determine tongue coating microbiome,” Sci. Rep., vol. 2, no. 1, p. 936, Dec. 2012.
[11] B. Pang, D. Zhang, N. Li, and K. Wang, “Computerized tongue diagnosis based on Bayesian networks,” IEEE Trans. Biomed. Eng., vol. 51, no. 10, pp. 1803–1810, Sep. 2004.
[12] Z. Li, Z. Yu, W. Liu, and Z. Zhang, “Tongue image segmentation via color decomposition and thresholding,” in Proc. 6th Int. Conf. Inf. Sci. Control Eng. (ICISCE), Jul. 2017, pp. 1–9.
[13] B. Pang, D. Zhang, and K. Wang, “The bi-elliptical deformable contour and its application to automated tongue segmentation in chinese medicine,” IEEE Trans. Med. Imag., vol. 24, no. 8, pp. 946–956, Aug. 2005.
[14] H. Zhang, W. Zuo, K. Wang, and D. Zhang, “A snake-based approach to automated segmentation of tongue image using polar edge detector,” Int. J. Imag. Syst. Technol., vol. 16, no. 4, pp. 103–112, 2007.
[15] J. Ning, D. Zhang, C. Wu, and F. Yue, “Automatic tongue image segmentation based on gradient vector flow and region merging,” Neural Comput. Appl., vol. 21, no. 8, pp. 1819–1826, Dec. 2010.
[16] M. Shi, G. Li, and F. Li, “CG2GFSnake: Automatic tongue image segmentation utilizing prior knowledge,” Sci. China Inf. Sci., vol. 56, no. 9, pp. 1–14, Dec. 2011.
[17] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins, “Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation,” NeuroImage, vol. 54, no. 2, pp. 940–954, Jan. 2011.
[18] G.-H. Liu and J.-Y. Yang, “Exploiting color volume and color difference for salient region detection,” IEEE Trans. Image Process., vol. 28, no. 1, pp. 6–16, Jan. 2019.
[19] C. Tian, Y. Xu, and W. Zuo, “Image denoising using deep CNN with batch renormalization,” Neural Netw., vol. 121, pp. 461–473, Jan. 2020.
[20] J. Wen, Y. Xu, Z. Li, Z. Ma, and Y. Xu, “Inter-class sparsity based discriminative least square regression,” Neural Netw., vol. 102, pp. 36–47, Jun. 2018.
[21] A. A. Efros and W. T. Freeman, “Image quilting for texture synthesis and transfer,” in Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn., 2001, pp. 341–346.

[22] M. J. Fadili, J.-L. Starck, and F. Murtagh, “Inpainting and zooming using sparse representations,” Comput. J., vol. 52, no. 1, pp. 64–79, Feb. 2008.

[23] L. Liu, L. Chen, C. L. P. Chen, Y. Y. Tang, and C. M. Pan, “Weighted joint sparse representation for removing mixed noise in image,” IEEE Trans. Cybern., vol. 47, no. 3, pp. 600–611, Mar. 2017.

[24] C. Tian, Y. Xu, Z. Li, W. Zuo, L. Fei, and H. Liu, “Attention-guided CNN for image denoising,” Neural Netw., vol. 124, pp. 117–129, Apr. 2020.

[25] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” IEEE Trans. Image Process., vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[26] Y. Gao, R. Ji, W. Liu, Q. Dai, and G. Hua, “Weakly supervised visual dictionary learning by harnessing image attributes,” IEEE Trans. Image Process., vol. 23, no. 12, pp. 5400–5411, Dec. 2014.

[27] L. Wei, S. Wan, J. Guo, and K. K. Wong, “A novel hierarchical selective ensemble classifier with bioinformatics application,” Artif. Intell. Med., vol. 83, pp. 82–90, Nov. 2017.

[28] Z. Liu, J. Wang, G. Liu, and L. Zhang, “Discriminative low-rank preserving projection for dimensionality reduction,” Appl. Soft Comput., vol. 85, Dec. 2019, Art. no. 105768.

[29] L. Wei, R. Su, B. Wang, X. Li, Q. Zou, and X. Gao, “Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites,” Neurocomputing, vol. 324, pp. 5–9, Jan. 2019.

[30] L. Wang, F. Shi, G. Li, Y. Gao, W. Lin, J. H. Gilmore, and D. Shen, “Segmentation of neonatal brain MR images using patch-driven level sets,” NeuroImage, vol. 84, pp. 141–158, Jan. 2014.

[31] A. Li, Z. Lu, L. Wang, P. Han, and J.-R. Wen, “Large-scale sparse learning from noisy tags for semantic segmentation,” IEEE Trans. Cybern., vol. 48, no. 1, pp. 253–263, Jan. 2018.

[32] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, “ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides,” Bioinformatics, vol. 34, no. 23, pp. 4007–4016, Jun. 2018.

[33] G. Huo, S. X. Yang, Q. Li, and Y. Zhou, “A robust and fast method for sidescan sonar image segmentation using nonlocal despckling and active contour model,” IEEE Trans. Cybern., vol. 47, no. 4, pp. 855–872, Apr. 2017.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[35] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” J. Roy. Stat. Soc. B Stat. Methodol., vol. 67, no. 2, pp. 301–320, Apr. 2005.

[36] R. Tibshirani, “Regression shrinkage and selection via the lasso,” J. Roy. Stat. Soc. B Methodol., vol. 58, no. 1, pp. 267–288, Dec. 2001.

[37] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” Ann. Stat., vol. 32, no. 2, pp. 407–499, 2004.

[38] H. Cheng, Z. Liu, and J. Yang, “Sparsity induced similarity measure for label propagation,” in Proc. IEEE 12th Int. Conf. Comput. Vis., Sep. 2009, pp. 317–324.

[39] W. A. Yasnoff, J. K. Mui, and J. W. Bacus, “Error measures for scene segmentation,” Pattern Recognit., vol. 9, no. 4, pp. 217–231, Jan. 1977.

[40] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, Jun. 2006.

[41] J. L. Fleiss, J. Cohen, and B. S. Everitt, “Large sample standard errors of kappa and weighted kappa,” Psychol. Bull., vol. 72, no. 5, pp. 323–327, 1969.

[42] X. Wang and D. Zhang, “A high quality color imaging system for computerized tongue image analysis,” Expert Syst. Appl., vol. 40, no. 15, pp. 5854–5866, Nov. 2013.

WEIXIA LIU received the B.S. degree in applied mathematics from Zhengzhou University, Zhengzhou, China, in 2007, and the M.S. degree in applied mathematics from Xi’an Jiaotong University, Xi’an, China, in 2010. She is currently a Lecturer with the Stratis Institute, Minjiang University, Fuzhou, China. She has published some articles in international/national journals. Her current research interests include image processing, pattern recognition, and statistical analysis.

ZHONGYI HU (Senior Member, IEEE) was born in Zhejiang, China. He received the B.S. degree in automation from the Beijing Institute of Mechanical Industry, in 1995, and the M.S. and Ph.D. degrees in control engineering and intelligent traffic engineering from the Wuhan University of Technology, in 2006 and 2015, respectively. He is currently an Associate Professor with the Intelligent Information Systems Institute, Wenzhou University. His current research interests include machine learning tracking and image processing and analysis.

CHANGEN ZHOU received the B.S. degree from the School of Mathematics, Henan Normal University, Xinxiang, China, in 2004, and the M.S. degree in computer science and technology from Fuzhou University, Fuzhou, China, in 2007. He is currently a Lecturer with the College of Traditional Chinese Medicine, Fujian University of Traditional Chinese Medicine, Fuzhou. His current research interests include image processing and machine learning.

ZUOYONG LI received the B.S. and M.S. degrees in computer science and technology from Fuzhou University, Fuzhou, China, in 2002 and 2006, respectively, and the Ph.D. degree from the School of Computer Science and Technology, Nanjing University of Science and Technology (NUST), Nanjing, China, in 2010. He is currently a Professor with the College of Computer and Control Engineering, Minjiang University, Fuzhou. He has published over 70 articles in international/national journals. His current research interests are image processing, pattern recognition, and machine learning.