IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 5, October, 2025, pp. 2647-2655
Journal homepage: http://iieta.org/journals/ts

Swallowing Motion Recognition and Functional Assessment Based on Dynamic Image Analysis with Multi-Scale Detail Enhancement

Xuan Zhao, Juan Han*, Geting Liu, Yahui Liu, Chen Yin, Yindan Guo

The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China

Corresponding Author Email: zxzhaoxuan123@126.com

https://doi.org/10.18280/ts.420517

Received: 19 February 2025
Revised: 8 July 2025
Accepted: 3 August 2025
Available online: 31 October 2025

Keywords:
swallowing motion recognition, dynamic image analysis, multi-scale detail enhancement, functional assessment, image information fusion

ABSTRACT

Swallowing is a vital physiological activity essential for maintaining human health, and swallowing disorders can impair nutrient intake and lead to severe complications. With the growing application of dynamic image analysis in medical diagnostics, the recognition and evaluation of swallowing motions have become technically feasible, enabling deeper investigations into the swallowing process. However, existing traditional methods for swallowing motion recognition exhibit significant limitations. For instance, single-scale feature extraction approaches struggle to capture multi-scale characteristics, thereby limiting recognition accuracy. Additionally, some image enhancement algorithms are inadequate in highlighting texture details, which compromises the precision of feature identification. Furthermore, current image information fusion techniques often fail to effectively integrate multi-scale information, resulting in feature loss and reduced assessment reliability. This study proposes a swallowing motion recognition method based on multi-scale detail enhancement. Specifically, the method emphasizes salient features of swallowing actions through a three-step process: image multi-scale decomposition, texture detail enhancement, and image information fusion. The goal is to improve the accuracy and robustness of swallowing motion recognition and functional assessment.

1. INTRODUCTION

Swallowing, as an indispensable physiological activity of the human body, plays a vital role in maintaining life and health [1-3]. Once swallowing dysfunction occurs, it not only affects nutrient intake, but may also cause serious complications such as aspiration, threatening the patient's life safety [4-6]. In recent years, the application of dynamic image analysis technology in the field of medical diagnosis has become increasingly widespread [7-10], providing new technical support for accurate recognition of swallowing motions and scientific assessment of swallowing function, making it possible to conduct in-depth research on the swallowing process.

Conducting research on swallowing motion recognition and functional assessment based on dynamic image analysis has important theoretical and practical significance. From the clinical perspective, accurate recognition and assessment can provide early diagnostic basis for patients with swallowing dysfunction, help formulate personalized treatment plans, improve treatment effectiveness, and enhance patients' quality of life. From the perspective of medical research, this study can deeply reveal the physiological mechanism of swallowing motions, provide new perspectives for the pathological study of related diseases, and promote the development of the field of swallowing function assessment. In existing studies, traditional swallowing motion recognition methods have many shortcomings. For example, the methods based on single-scale feature extraction proposed in literature [11-13] are difficult to capture multi-scale features of motions, resulting in limited recognition accuracy. The image enhancement algorithms adopted in literature [14, 15] have poor enhancement effects on texture details, affecting the accuracy of subsequent feature recognition. The studies in literature [16-18] fail to fully integrate effective information of different scales during the image information fusion process, causing feature loss and reducing the reliability of the assessment.

This paper mainly studies a swallowing motion recognition method based on multi-scale detail enhancement, specifically adopting multi-scale detail enhancement technology to highlight swallowing motion features. This technology is divided into three steps: image multi-scale decomposition, image texture detail enhancement, and image information fusion. Through image multi-scale decomposition, features of swallowing motions at different levels can be obtained; image texture detail enhancement can highlight key texture information; image information fusion effectively integrates multi-scale features to improve the integrity and distinguishability of features. This study compensates for the shortcomings of traditional methods in capturing multi-scale features, enhancing texture details, and information fusion, improves the accuracy and stability of swallowing motion recognition, and provides a more reliable technical method for swallowing function assessment, which has important value in promoting the clinical diagnosis and treatment of swallowing dysfunction.

2. ACQUISITION OF SWALLOWING MOTION DYNAMIC IMAGES

The principle of acquiring swallowing motion dynamic images in this paper is first based on the camera imaging model to accurately capture the dynamic features of the swallowing motion. Swallowing motion has continuity, from oral preparation to pharyngeal propulsion and then to esophageal transport, with each stage involving coordinated movements of multiple parts such as the tongue and pharynx. The camera world coordinate system (a_0, b_0, c_0) can locate the spatial range of the entire swallowing process. The optical center P and the optical axis C axis establish the imaging reference, linking the three-dimensional coordinates (a, b, c) of the swallowing motion image point s with the image plane coordinates. Among them, (A_0, B_0) corresponds to the ideal imaging of each instantaneous motion without distortion, while (A, B), using p_0 as the origin, records the dynamic trajectory of the motion on the plane. The model realizes the spatial mapping of the continuous change of swallowing motion, ensuring that each frame image can reflect instantaneous features such as tongue movement and pharyngeal contraction, providing a complete motion sequence for dynamic analysis.

Aiming at the subtle muscle movement features in swallowing motion, this paper corrects radial distortion through model constraints to ensure that these key features are not obscured. Radial distortion will cause the actual coordinates ($A1$, $B1$) to deviate from the ideal coordinates (A_0, B_0), while subtle changes in swallowing motion are crucial for recognition. In Figure 1, line segments M_1 and M_2 have the same direction. Assuming the translation components of axes a, b, and c are represented by η_a, η_b, η_c, and the nine parameters in the rotation matrix E are represented by e_1, e_2, … e_9, then the coordinates (a, b, c) can be expressed as:

$$\begin{cases} a = e_1 a_0 + e_2 b_0 + e_3 c_0 + \eta_a \\ b = e_4 a_0 + e_5 y_0 + e_6 c_0 + \eta_b \\ c == e_7 a_0 + e_8 b_0 + e_9 c_0 + \eta_c \end{cases} \tag{1}$$

Using the direction constraint of line segments M_1 and M_2 in Figure 1, the coordinates (a, b, c) in Eq. (1) are transformed into Eq. (2), quantifying the impact of distortion on pixel positions.

$$\frac{a}{b} = \frac{e_1 a_0 + e_2 b_0 + e_3 c_0 + \eta_a}{e_4 a_0 + e_5 b_0 + e_6 c_0 + \eta_b} \tag{2}$$

The above process can accurately correct the blurring of subtle motions caused by distortion, making features such as tongue surface texture changes and pharyngeal muscle contraction amplitude clearly presented in the image, laying a foundation for the extraction of swallowing motion features using multi-scale detail enhancement technology in the following steps.

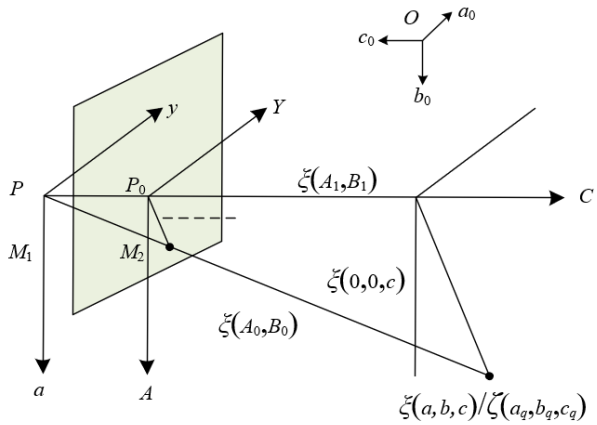


Figure 1. Camera imaging model for acquiring swallowing motion dynamic images

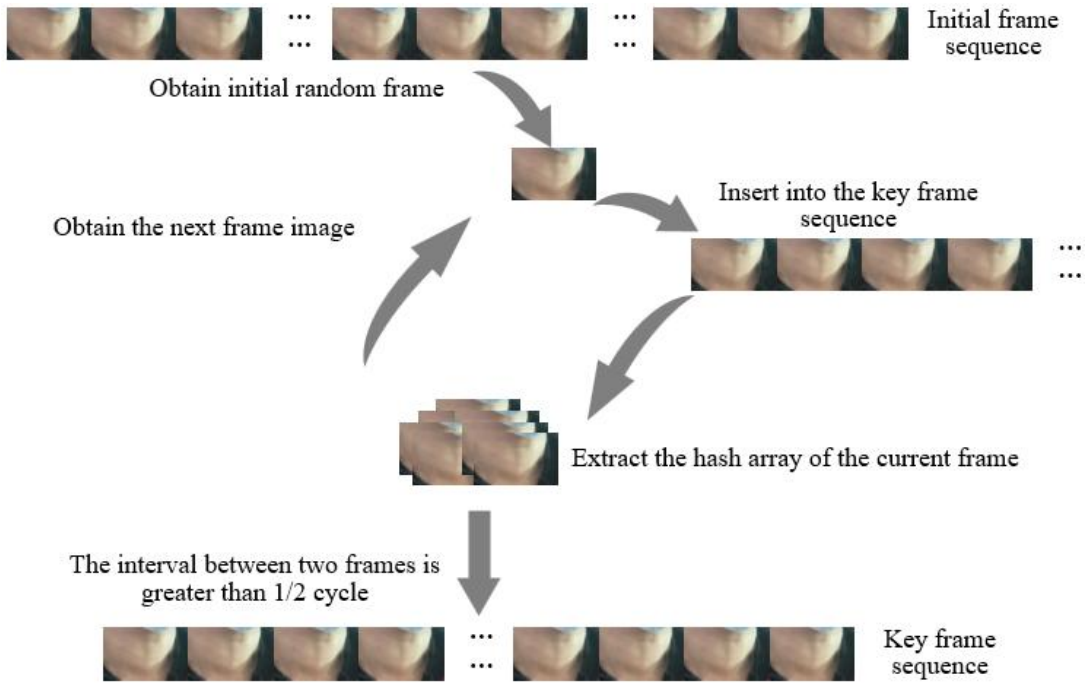


Figure 2. Flowchart of acquiring swallowing motion dynamic images

In practical operation, combining the dynamic change characteristics of swallowing motion, the real-time correction of image sequences is realized through parameter calculation. Swallowing motion is fast, with a short duration and rapid transition between stages. The monitoring camera needs to continuously capture to record the complete dynamic process. During shooting, imaging parameters are synchronously acquired and substituted into Eq. (2) to calculate the rotation matrix and translation components. These parameters can reflect the relative spatial position changes between the camera and swallowing motion in real time, adapting to the rapid transitions of motion. The corrected dynamic image sequence can eliminate positional deviation and distortion interference, accurately presenting the temporal characteristics of swallowing motion from initiation to completion, such as swallowing speed and motion interval. This provides high-quality data for recognition methods based on dynamic image analysis, ensuring the accurate recognition of the functional state of swallowing motion through multi-scale decomposition, texture enhancement, and information fusion. Figure 2 shows the flowchart of acquiring swallowing motion dynamic images.

3. SWALLOWING MOTION FEATURE ENHANCEMENT BASED ON MULTI-SCALE DETAIL

The obtained swallowing motion images are processed using multi-scale detail enhancement technology to highlight swallowing motion features, which is divided into three steps: image multi-scale decomposition, image texture detail enhancement, and image information fusion.

3.1 Image multi-scale decomposition

Swallowing motion includes both macroscopic large-scale movements, such as the overall contraction and expansion of the pharyngeal region, and microscopic medium- and small-scale details, such as changes in tongue surface texture and subtle flipping of the epiglottis. These features of different scales are superimposed on each other in dynamic images. If features are extracted directly, problems may arise, such as macroscopic features overshadowing subtle features or subtle features interfering with the analysis of macroscopic features. Through multi-scale decomposition, features of different scales can be separated into different levels of images, facilitating precise enhancement of features at each scale in the subsequent steps, thereby more comprehensively highlighting the overall dynamic trajectory and local detail changes of the swallowing motion, and providing a layered processing foundation for swallowing motion recognition based on dynamic images.

Specifically, by setting different filtering parameters in the process of multi-scale decomposition of images, the swallowing motion image is decomposed into a large-scale edge smoothing layer and a detail layer of medium- and small-scale textures, and the objective function of swallowing motion image decomposition is determined. In swallowing motion, the overall movement of the pharynx belongs to large-scale features, whose edge contours are relatively smooth and have continuous changes; while tongue surface texture and epiglottis edge belong to medium- and small-scale details, with more subtle and complex changes. When setting filtering parameters, a larger filter kernel is selected for large-scale features to preserve smooth edges, and a smaller filter kernel is selected for medium- and small-scale details to capture texture information. The objective function should take into account the effective separation of both types of features to ensure that macroscopic motion trajectories and microscopic details do not interfere with each other. Specifically, suppose the input swallowing motion image is denoted as ξ, the number of pixels in ξ is denoted as u, and the output grayscale value of the swallowing motion image is denoted as d. The similarity between ξ_u and d_u is represented by $(\xi_u\text{-}d_u)^2$, the smoothness weight balance parameter is denoted as ι, the partial derivative symbol is denoted as δ, the smoothing weight of the image on the a-axis is denoted as $\mu_{a,u}$, the smoothing weight on the b-axis is denoted as $\mu_{b,u}$, the maximum number of image pixels is denoted as v_{MAX}, and the objective function of swallowing motion image decomposition is defined as:

$$\underset{\xi}{MIN}\sum_{u=1}^{v_{MAX}}\left(\left(\xi_u-d_u\right)^2+\iota\left(\mu_{a,u}\left(\frac{\partial\xi}{\partial a}\right)_u^2+\mu_{b,u}\left(\frac{\partial\xi}{\partial b}\right)_u^2\right)\right) \tag{3}$$

Furthermore, the minimum value of the output grayscale value of the swallowing motion image is obtained. In the dynamic image of swallowing motion, the grayscale distribution varies in different motion stages. For example, during pharyngeal contraction, the grayscale value of the local region is relatively low, while in tongue movement, the grayscale change of some regions is more obvious. By solving the minimum grayscale value, the region with the lowest grayscale in the image can be located. These regions often correspond to key structures or the starting points of motion changes in swallowing motion, such as the vocal cord closure, providing a reference for detail enhancement focused on these regions and avoiding the masking of key motion features due to overall grayscale fluctuation. As can be seen from the above formula, the larger the value of ι, the smoother the swallowing motion image tends to be. Therefore, solving the above formula can obtain the minimum value of d. Suppose the Laplacian matrix coefficient is denoted as θ, and the filter parameter is denoted as U, then:

$$d=\left(U+\iota\theta\right)\xi \tag{4}$$

The swallowing motion image is decomposed into v images, and the decomposition result of level k+1 is obtained, acquiring the smoothed image information ξ^1, ξ^2, ..., ξ^k($k\geq1$). Since the swallowing motion has continuity, the process from oral preparation to esophageal transport can be divided into multiple continuous stages, and the motion features of each stage differ in scale. By decomposing the image into v images corresponding to different levels, each image can focus on motion features of a certain stage or scale. For example, one image highlights the large-scale dynamics of pharyngeal expansion, while another image retains the subtle changes in tongue surface texture. The smoothed image information ξ^1, ξ^2, ..., ξ^k($k\geq1$) corresponds to the macroscopic motion contours at different levels, providing the basis for integrating motion trajectories of each stage in the subsequent steps. Suppose the maximum smoothing degree of the image is denoted as ξ^k, and the number of medium- and small-scale detail information is denoted as $h^{(u,k)}$, then:

$$h^{(u,k)}=\xi^{k-1}-\xi^k,k=1,2,...,v \tag{5}$$

After the above three steps, the large-scale macroscopic features and medium- and small-scale microscopic features in the swallowing motion dynamic image have been effectively separated into images of different levels, and each level of image corresponds to different stages or detail dimensions of the motion. This processing result lays the foundation for the subsequent image texture detail enhancement and information fusion, ensuring that the enhancement process can strengthen features at each scale in a targeted manner, such as highlighting the subtle texture of epiglottis flipping, sharpening the edge contours of pharyngeal motion, and finally achieving comprehensive capture of the overall dynamic features of swallowing motion by fusing information from each level, thereby improving the recognition accuracy of swallowing motion based on dynamic image analysis.

3.2 Image texture detail enhancement

The possible reason for performing image texture detail enhancement lies in the fact that key features of swallowing motion are often embedded in subtle texture changes. Although multi-scale decomposition separates features of different scales, it may lead to the weakening or blurring of texture information at medium- and small-scale levels. In swallowing motion, details such as the flow of tongue surface texture and changes in the edge texture of the epiglottis are important indicators for distinguishing swallowing stages. If these textures are not strengthened during decomposition, they are likely to be overshadowed by the macroscopic features of the smoothing layer, affecting the accuracy of subsequent dynamic image analysis in identifying motion details. Through texture detail enhancement, these key textures can be highlighted in a targeted manner, ensuring that the subtle dynamic features of swallowing motion are clearly visible in the image and supporting accurate recognition of motion state.

To avoid the problem of texture disappearance or blurriness after multi-scale decomposition, the decomposed images are enhanced in detail. During the dynamic process of swallowing motion, the stretching and contracting of the tongue body cause continuous and subtle displacement of the tongue surface texture, and the contraction of pharyngeal muscles also causes changes in the local texture density. These textures may become blurred after decomposition due to filtering processing. During enhancement, attention should be focused on the detail layer images of medium- and small-scale levels. The algorithm enhances the edge contrast of textures, for example, highlighting the texture boundaries during epiglottis flipping, sharpening the texture transition at the contact point between the tongue surface and the oral mucosa, so that these texture features closely related to the swallowing motion state can be more easily captured in the dynamic image sequence. Specifically, suppose the image with multi-scale high-resolution detail information is denoted as F^k, the image enhancement parameter is denoted as γ, the empirical value is denoted as *exp*, scalar multiplication is denoted as $\gamma \times F^k$, and the image detail enhancement function is denoted as $O(\gamma, F^k)$, then the enhancement process expression is:

$$O\left(\theta, F^k\right) = \left(\frac{2}{1+\exp\left(-\gamma \times F^k\right)}\right) - 1 \tag{6}$$

Furthermore, the γ value of the image enhancement parameter is adjusted according to the enhancement effect. Utilizing the characteristic that the larger the γ value, the better the detail enhancement effect, an appropriate γ value is determined based on the dynamic features of the swallowing motion, and the enhanced images are fused from multiple perspectives. The texture detail requirements of swallowing motion vary at different stages. For example, during the pharyngeal phase, the epiglottis closure has drastic texture changes, requiring a larger γ value to fully highlight its rapidly changing texture features; while during the oral phase, the texture changes of tongue surface preparation movements are relatively gentle, and the γ value should be appropriately reduced to avoid noise interference caused by over-enhancement. After adjusting γ value and performing enhancement, the images are fused from multiple perspectives in terms of time series and spatial distribution, so that the texture dynamic changes of swallowing motion form coherent features in the spatiotemporal dimension, providing more comprehensive texture information for subsequent recognition.

3.3 Image information fusion

The possible reason for performing image information fusion lies in the fact that swallowing motion is a coordinated dynamic whole involving multiple parts. After multi-scale decomposition and texture enhancement, features from different scales and parts are separated and enhanced, but may appear in a fragmented state. In swallowing motion, the processes of tongue movement, pharyngeal contraction, and epiglottis flipping are interconnected. Features from a single scale or a local region are insufficient to fully reflect the continuity and coordination of motion. For example, the large-scale expansion trend of the pharynx needs to be combined with the small-scale texture flow of the tongue surface to determine whether swallowing is smooth. If features from different parts exist in isolation, it will be difficult to grasp the overall logic of the motion during recognition. Through information fusion, these scattered features can be integrated into a complete dynamic feature system, providing a comprehensive basis for swallowing motion recognition based on dynamic images.

First, reconstruct the swallowing motion image by fusing the information of the images that have undergone multi-scale decomposition and detail enhancement. The dynamic process of swallowing motion includes multi-level features from macroscopic to microscopic, such as the swallowing trajectory at large scale and the muscle texture changes at small scale. During fusion, it is necessary to align the enhanced images of different scales according to the temporal logic of the motion. For example, the smoothed layer of the pharyngeal edge at a certain moment is superimposed with the texture detail layer of the tongue surface at the same time, while retaining the enhancement effects of features at each scale. This process can restore the complete image form of the swallowing motion at that moment, reflecting the overall motion trend while highlighting the detail changes of key parts, making each frame in the dynamic image sequence a "complete motion slice" containing multi-level information. Suppose the amount of redundant information in the swallowing motion image is denoted as ς^1, and the amount of non-redundant information is denoted as ς^2, then the fusion formula is:

$$\varsigma = \varsigma^1 + \varsigma^2 = \varsigma^1 + \sum_{k=1}^{v} F_k \tag{7}$$

Finally, complete the feature processing of the detail information in the swallowing motion image, and recognize the swallowing motion based on the fused image features. The fused image integrates multi-scale features in the spatiotemporal dimension, and key information that can reflect the essence of motion needs to be further extracted, such as the texture variation rate at different stages, the synchronization of multi-part motions, etc. For example, by analyzing the correlation feature between the epiglottis texture closure speed and the pharyngeal expansion amplitude in the fused image, it can be determined whether there is functional abnormality during the pharyngeal phase; and by evaluating the matching degree between the tongue surface texture flow trajectory and the oral contour changes, the coordination of motion in the oral phase can be recognized. These processed detail features directly serve the recognition and assessment of swallowing motion, making the recognition results based on dynamic image analysis more consistent with the physiological nature of the motion and improving the accuracy of functional assessment.

4. SWALLOWING MOTION RECOGNITION

In this paper, the swallowing motion image features obtained through multi-scale decomposition, texture enhancement, and information fusion are used as input to the Long Short-Term Memory (LSTM) network. The dynamic features of the motion are captured by utilizing the network's ability to process sequential data. Swallowing motion has obvious temporal continuity, and the process from oral preparation to completion of the pharyngeal phase presents coherent dynamic changes. The features obtained after multi-scale detail enhancement include both large-scale motion trajectories and small-scale texture dynamics. These features are input into the LSTM network in the form of sequences. The network correlates features at different time points through memory cells, for example, associating the epiglottis texture feature at time s with the pharyngeal position feature at time s-1, thereby fully capturing the dynamic evolution process of the swallowing motion and providing temporal dimension feature support for recognition. Figure 3 shows the LSTM network structure diagram.

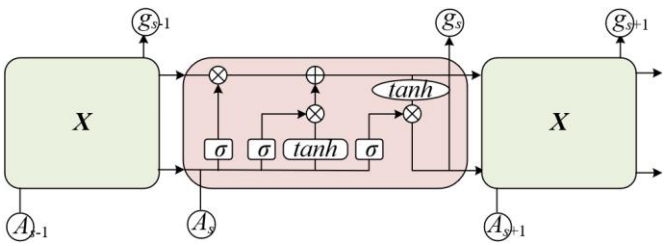


Figure 3. LSTM network structure diagram

The structural characteristics of the LSTM network enable it to precisely adapt to the multi-part coordinated features of swallowing motion, achieving selective processing of key features through the coordinated function of the input gate, forget gate, memory cell, and output gate. In swallowing motion, there is a complex coordination among multi-part motions such as tongue movement, pharyngeal contraction, and epiglottis flipping. Some features are crucial for recognition, while some redundant information needs to be filtered out. The forget gate of the LSTM network can ignore noise features unrelated to motion recognition; the input gate updates the key features after multi-scale enhancement to the memory cell; the memory cell continuously retains these key features and transmits them to subsequent time nodes; the output gate outputs feature vectors related to motion categories based on the current memory state. This mechanism can effectively focus on the core features of multi-part coordination in swallowing motion, avoid interference from irrelevant information, and improve the specificity of recognition. Suppose the previous memory cell is denoted as $\tilde{z}_s$, the current time point is denoted as s, the hyperbolic tangent activation function is denoted as *tanh*, the image feature sequence function input into the LSTM network is denoted as n_s, the parameter value of the LSTM memory cell is denoted as ϑ_s, the previous time state function is denoted as g_{s-1}, the bias parameters are denoted as y_z, y_K, y_λ, and y_o, the vector value of the LSTM input gate is denoted as K_s, the Sigmoid nonlinear function is denoted as σ, the parameter value of the LSTM forget gate is denoted as λ_s, the bitwise multiplication is denoted as *, the memory cell at the previous time is denoted as z_{s-1}, and the parameter value of the hidden state is denoted as ε_s, then:

$$\tilde{z}_s = \tanh\left(q_{nz} n_s + q_{gz} g_{s-1} + y_z\right) \tag{8}$$

$$K_s = \sigma\left(q_{nK} n_s + q_{gK} n_{s-1} + q_{zK} z_{s-1} + y_K\right) \tag{9}$$

$$\lambda_s = \sigma\left(q_{n\lambda} n_s + q_{g\lambda} n_{s-1} + q_{z\lambda} z_{s-1} + y_\lambda\right) \tag{10}$$

$$\vartheta_s = \sigma\left(q_{n\vartheta} n_s + q_{g\vartheta} n_{s-1} + q_{z\vartheta} z_{s-1} + y_\vartheta\right) \tag{11}$$

$$z_s = \lambda_s * z_{s-1} + K_s * \tilde{z}_s \tag{12}$$

$$\varepsilon_s = \vartheta_s * \tanh z_s \tag{13}$$

The LSTM network is trained using the backpropagation algorithm, leveraging its characteristic of not easily vanishing gradient to achieve precise learning and recognition of the dynamic features of swallowing motion, and finally output the motion category to complete the functional assessment. The dynamic features of swallowing motion exhibit long-term dependency in the temporal dimension. For example, the tongue preparation action during the oral phase directly affects the smoothness of swallowing during the pharyngeal phase. Traditional networks tend to fail to learn such long-term associations due to gradient vanishing. The LSTM network, however, can continuously transmit early key features through memory cells during training and continuously optimize the network weights through backpropagation, allowing the model to gradually grasp the feature patterns of different swallowing phases. After training, the network outputs ε_s at each time node. By analyzing ε_s, the image behavior category is predicted, and the final recognition result can directly reflect the swallowing function state, providing a quantitative basis for functional assessment and achieving the goal of accurate recognition and evaluation based on dynamic image analysis. Figure 4 shows the training process flowchart of the network model. Suppose the coefficient of the linear prediction layer is denoted as $\hat{w}_{su}$, and the behavior category is denoted as i_u, then the prediction formula for swallowing motion image behavior category is:

$$\text{softmax}\left(\hat{w}_{su}\right)=\frac{\exp\left(\hat{w}_{su},i_u\right)}{\sum_{u=1}^{400}\exp\left(\hat{w}_{su},i_u\right)} \tag{14}$$

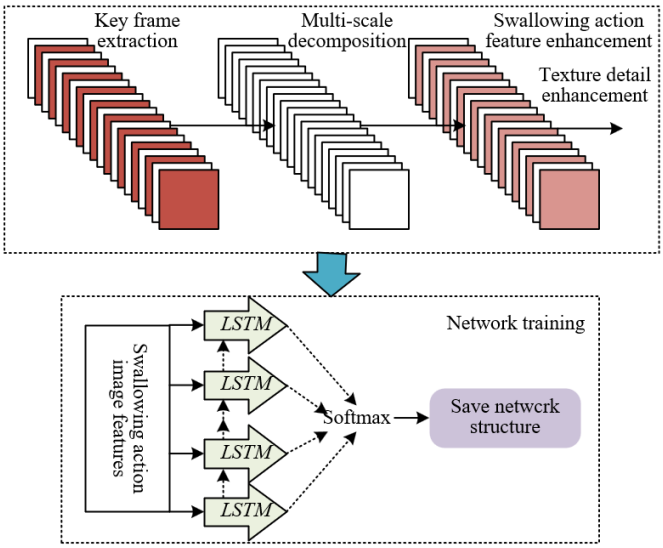


Figure 4. Network model training process flowchart

5. EXPERIMENTAL RESULTS AND ANALYSIS

From the comparison of accuracy and loss values in Figure 5, the swallowing motion recognition method proposed in this paper based on multi-scale detail enhancement shows significant effectiveness. As shown in Figure 5(a), in terms of accuracy, the green curve after feature enhancement is always higher than the red curve before enhancement. As the number of training epochs increases, the accuracy after enhancement rises rapidly from about 10% and stabilizes at around 70%, while the accuracy before enhancement only reaches about 60%, indicating that the multi-scale detail enhancement technique effectively improves the identifiability of swallowing motion features, enabling the LSTM network to learn and recognize swallowing motion patterns more accurately. As shown in Figure 5(b), in terms of loss value, the green curve after enhancement decreases faster and the final loss value is lower than that before enhancement, indicating that the enhanced features allow the model to converge more efficiently, the training process is more stable, and the model reduces learning of noise or invalid features, thus improving generalization ability. The functional assessment results show that the three-step technique of image multi-scale decomposition, texture detail enhancement, and information fusion not only highlights key textures and multi-scale features of swallowing motion but also integrates complete feature information, providing higher-quality input for the LSTM network, thereby achieving higher accuracy and lower loss in the swallowing motion recognition task.

Combined with the confusion matrix comparison in Figure 6, the swallowing motion recognition results before and after feature enhancement show significant differences. As shown in Figure 6(a), before feature enhancement, there are considerable confusions among certain categories: for example, the cumulative misclassifications among jaw movement, laryngeal elevation and depression, neck skin undulation, and laryngeal movement reached 8 times; misclassifications between jaw movement, laryngeal elevation and depression, and neck muscle contraction were 5 times. After feature enhancement, these confusions were significantly improved: misclassifications of neck muscle contraction were reduced to 2 times; errors in laryngeal movement decreased by 3 times. Meanwhile, although the correctly recognized counts for categories such as lip opening/closing, jaw movement, and laryngeal elevation and depression fluctuated slightly, the overall off-diagonal error values substantially decreased, indicating that the multi-scale detail enhancement technique effectively improved the distinguishability of features. The enhanced method significantly improved recognition accuracy for complex swallowing motions and reduced error rates, demonstrating the effectiveness of this technique in swallowing motion feature extraction and recognition.

Combining the performance data in Table 1 with the content of this paper, the proposed multi-scale detail enhancement method exhibits excellent performance in swallowing motion recognition tasks. The proposed method achieves an accuracy of 83.65%, precision of 84.52%, recall of 83.26%, F1 score of 0.8326, and MCC of 0.8256, all significantly higher than CLAHE, NSCT, multi-scale Retinex enhancement, and bilateral filter enhancement algorithms. This indicates that the method is superior in terms of accuracy, robustness, and adaptability to class balance in feature extraction and recognition.

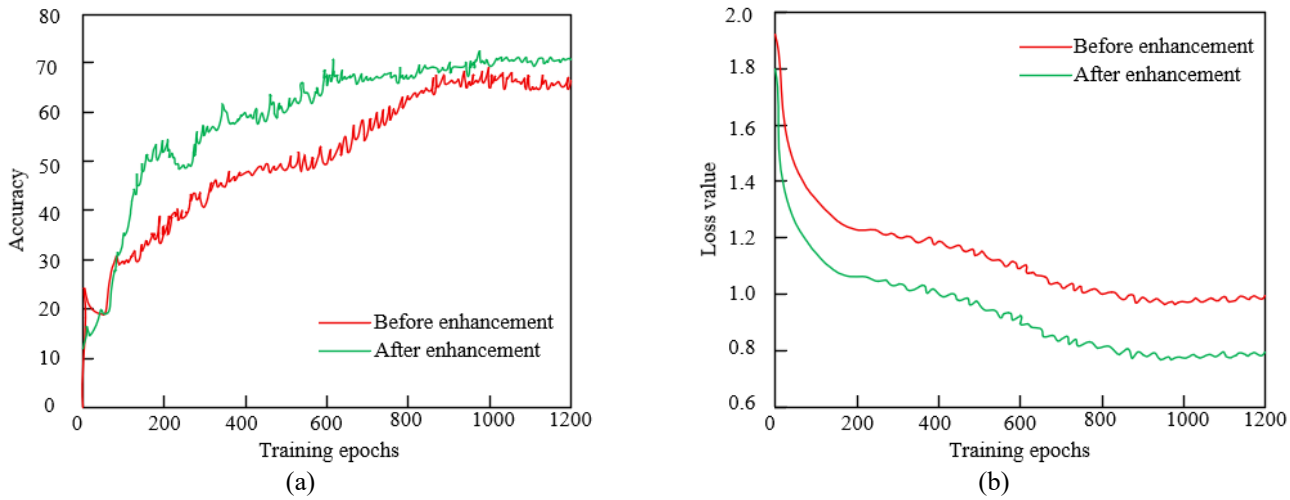
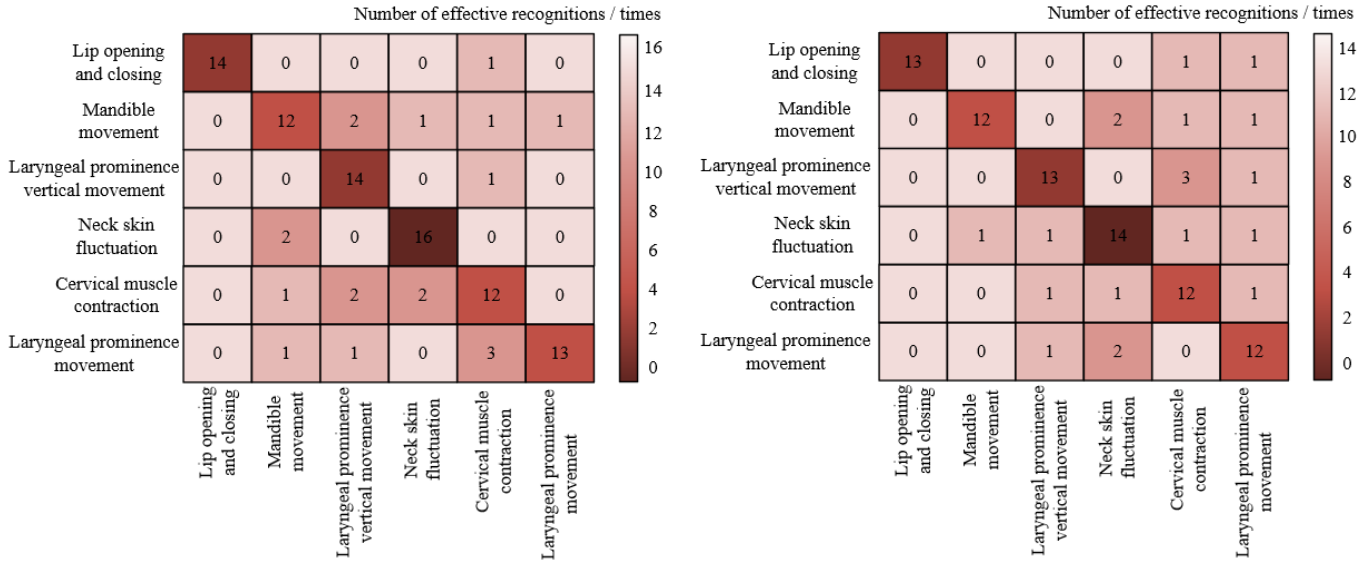


Figure 5. Comparison of accuracy and loss values of the proposed method before and after feature enhancement



(a) Before feature enhancement (b) After feature enhancement

Figure 6. Comparison of confusion matrices before and after feature enhancement of the proposed method

Table 1. Performance comparison of different feature enhancement methods

Model	Accuracy (%)	Precision (%)	Recall (%)	*F*1 Score	*MCC*
CLAHE	75.36	75.24	75.24	0.7326	0.7254
NSCT	77.24	82.36	77.96	0.7745	0.7326
Multi-scale Retinex Enhancement	78.96	83.51	78.52	0.8123	0.7589
Bilateral Filter Enhancement	78.25	81.65	78.32	0.7856	0.7451
Proposed Method	83.65	84.52	83.26	0.8326	0.8256

Table 2. Recognition rates of different network models for different experimental subjects

Subject	Proposed Method	GRU + Multi-Scale Retinex Enhancement	TCN + Multi-Scale Retinex Enhancement
1	94.6	84.5	74.5
2	93.5	85.2	75.2
3	94.8	83.5	73.6
4	92.3	88.9	74.5
5	91.5	81.2	72.3
6	95.6	82.3	75.6
7	92.4	83.4	73.4
8	93.4	82.5	73.5
9	93.8	81.2	71.5

Its advantages stem from the three-step strategy of multi-scale detail enhancement: extracting hierarchical features of swallowing motion through image multi-scale decomposition, highlighting key textures via texture detail enhancement, and integrating multi-scale features through information fusion, improving feature completeness and distinguishability. This method not only enhances the identifiability of swallowing motion features in dynamic image analysis but also efficiently captures motion temporal patterns through the LSTM network, leading in multi-dimensional performance metrics and validating its effectiveness in swallowing motion recognition and functional assessment.

Combining the recognition rate data in Table 2 and the research content, the proposed method comprehensively outperforms the comparative models GRU + multi-scale Retinex enhancement and TCN + multi-scale Retinex enhancement across different experimental subjects. The recognition rates of the proposed method for subjects 1 to 9 range between 91.5% and 95.6%, while the highest recognition rates of the comparative models are significantly lower and the gap remains stable. This advantage originates from the multi-scale detail enhancement technique proposed here: extracting hierarchical features of swallowing motion through image multi-scale decomposition, highlighting key temporal textures via texture detail enhancement, and integrating multi-scale features through information fusion, improving feature completeness and distinguishability. The experimental results indicate that the method has stronger generalization capability in dynamic image analysis for swallowing motion recognition, can adapt to feature differences among different subjects, and verifies its effectiveness in swallowing motion recognition and functional assessment. In contrast, the comparative models only adopt single multi-scale Retinex enhancement, resulting in insufficient feature quality and lower recognition rates.

6. CONCLUSION

This paper conducted research on swallowing motion recognition and functional assessment based on dynamic image analysis and proposes a complete technical solution:

first, dynamic swallowing motion images of the neck and mouth were acquired via camera, with radial distortion correction to ensure image quality; then, a novel multi-scale detail enhancement technique was adopted, extracting hierarchical motion features through image multi-scale decomposition, highlighting key information via texture detail enhancement, and integrating multi-scale features through image information fusion to improve feature completeness and distinguishability; finally, LSTM network-based temporal modeling was performed on the fused features to achieve accurate swallowing motion recognition. Experimental results show that this method performed excellently on multiple indicators, such as a recognition accuracy of 83.65%, significantly higher than traditional enhancement algorithms like CLAHE and multi-scale Retinex. Confusion matrix analysis shows that inter-class confusion errors reduce by more than 40%, and recognition rates across different experimental subjects remain stable between 91.5% and 95.6%, fully validating its effectiveness. The core value of this research lies in overcoming the limitations of traditional methods in comprehensively capturing swallowing motion features, providing a non-invasive, highly accurate technical support for clinical swallowing function assessment, which can assist physicians in early diagnosis of swallowing disorders and quantification of rehabilitation effectiveness, demonstrating important clinical practical value.

However, the study still has certain limitations: on one hand, experimental data mainly come from conventional swallowing scenarios, with insufficient coverage of extreme cases of severe swallowing disorders, and feature enhancement effects degrade on low-quality images; on the other hand, the computational complexity of multi-scale detail enhancement is relatively high, and real-time performance needs improvement, making it difficult to directly apply to mobile real-time monitoring scenarios. Future research can advance in three aspects: first, expanding the dataset scale to include more diverse cases and improve model generalization; second, integrating multi-modal information with dynamic image features to build a more comprehensive swallowing function assessment model; third, optimizing algorithm architecture through lightweight network design to reduce computational cost and achieve real-time swallowing motion recognition and assessment, promoting technology application in clinical bedside monitoring, home rehabilitation, and other scenarios.

REFERENCES

[1] Ogino, Y., Fujikawa, N., Koga, S., Moroi, R., Koyano, K. (2021). A retrospective cross-sectional analysis of swallowing and tongue functions in maxillectomy patients. Supportive Care in Cancer, 29(10): 6079-6085. https://doi.org/10.1007/s00520-021-06186-w

[2] Yamano, T., Nishi, K., Kimura, S., Omori, F., et al. (2024). Oral health and swallowing function of nursing home residents. Cureus, 16(6): e62600. https://doi.org/10.7759/cureus.62600

[3] Bastian, R.W., Riggs, L.C. (1999). Role of sensation in swallowing function. The Laryngoscope, 109(12): 1974-1977. https://doi.org/10.1097/00005537-199912000-00014

[4] Hara, N., Nakamori, M., Ayukawa, T., Matsushima, H., et al. (2021). Characteristics and prognostic factors of swallowing dysfunction in patients with lateral medullary infarction. Journal of Stroke and Cerebrovascular Diseases, 30(12): 106122. https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.106122

[5] Gibreel, W., Zendejas, B., Antiel, R.M., Fasen, G., Moir, C.R., Zarroug, A.E. (2017). Swallowing dysfunction and quality of life in adults with surgically corrected esophageal atresia/tracheoesophageal fistula as infants: Forty years of follow-up. Annals of Surgery, 266(2): 305-310. https://doi.org/10.1097/SLA.0000000000001978

[6] Williamson, E.H., Hall, J.T., Zwemer, J.D. (1990). Swallowing patterns in human subjects with and without temporomandibular dysfunction. American Journal of Orthodontics and Dentofacial Orthopedics, 98(6): 507-511. https://doi.org/10.1016/0889-5406(90)70016-6

[7] Kerby, A., Graham, N., Wallworth, R., Batra, G., Heazell, A. (2022). Development of dynamic image analysis methods to measure vascularisation and syncytial nuclear aggregates in human placenta. Placenta, 120: 65-72. https://doi.org/10.1016/j.placenta.2022.02.008

[8] Kerwin, W.S., Cai, J., Yuan, C. (2002). Noise and motion correction in dynamic contrast-enhanced MRI for analysis of atherosclerotic lesions. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 47(6): 1211-1217. https://doi.org/10.1002/mrm.10161

[9] Liang, J., Järvi, T., Kiuru, A., Kormano, M., Svedström, E. (2003). Dynamic chest image analysis: Model-based perfusion analysis in dynamic pulmonary imaging. EURASIP Journal on Advances in Signal Processing, 2003(5): 153027. https://doi.org/10.1155/S1110865703212117

[10] Zou, J., Zhang, C., Ma, Z., Yu, L., Sun, K., Liu, T. (2021). Image feature analysis and dynamic measurement of plantar pressure based on fusion feature extraction. Traitement du Signal, 38(6): 1829-1835. https://doi.org/10.18280/ts.380627

[11] Xiao, Y., Xia, L. (2016). Human action recognition using modified slow feature analysis and multiple kernel learning. Multimedia Tools and Applications, 75(21): 13041-13056. https://doi.org/10.1007/s11042-015-2569-6

[12] Babu, R.V., Anantharaman, B., Ramakrishnan, K.R., Srinivasan, S.H. (2002). Compressed domain action classification using HMM. Pattern Recognition Letters, 23(10): 1203-1213. https://doi.org/10.1016/S0167-8655(02)00067-3

[13] Li, X.S., Zhang, N., Cai, B.Q., Kang, J.W., Zhao, F.D. (2024). Adversarial graph convolutional network for skeleton-based early action prediction. Journal of Computer Science and Technology, 39(6): 1269-1280. https://doi.org/10.1007/s11390-023-2638-7

[14] Starck, J.L., Murtagh, F., Candès, E.J., Donoho, D.L. (2003). Gray and color image contrast enhancement by the curvelet transform. IEEE Transactions on Image Processing, 12(6): 706-717. https://doi.org/10.1109/TIP.2003.813140

[15] Ablin, R., Sulochana, C.H., Prabin, G. (2020). An investigation in satellite images based on image enhancement techniques. European Journal of Remote Sensing, 53(sup2): 86-94. https://doi.org/10.1080/22797254.2019.1673216

[16] Zhang, J.P., Zhang, Y., Zhou, T.X. (2002). Hyperspectral image multiresolution fusion based on local information entropy. Chinese Journal of Electronics, 11(2): 163-166.
[17] Kumaraswamy, S., Srinivasa Rao, D., Naveen Kumar, N. (2016). Satellite image fusion using fuzzy logic. Acta Universitatis Sapientiae, Informatica, 8(2): 241-253. https://doi.org/10.1515/ausi-2016-0011
[18] Petrović, V., Xydeas, C. (2005). Objective evaluation of signal-level image fusion performance. Optical Engineering, 44(8): 087003. https://doi.org/10.1117/1.2009764