Gesture identification in testing mode in the integrated sign language training system

M G Grif, A V Lukoyanychev

Novosibirsk State Technical University, 20, Karla Marksa ave., Novosibirsk, 630073, Russian Federation

E-mail: grifmg@mail.ru

Abstract. The structure of an integral sign language teaching system based on a 3D computer character, which allows demonstrating fingerprints and words of a sign language using an avatar, is considered. The system provides the test mode when the user must demonstrate the word or fingerprint indicated by the system. The user feedback with the system is carried out using a standard webcam. The problems of gesture localization and identification in user testing mode are considered. The features of this mode are determined during the system operation. The algorithm for quick identification of human hand gestures is proposed. The method based on the signs, which are two-dimensional projections, is used for gesture identification. The projection recognition features are considered, the similarity metrics are analyzed, and three stages for match decision are identified. The experimental results show the stable operation of this algorithm in various lighting conditions and at the place where the gesture was shown.

1. Introduction

The visual representation of 3D images is one of the methods to increase human efficiency information perception. The use of a computer character (avatar) to demonstrate sign language (SL) provides additional opportunities for further work in this direction. The developed integral system for teaching Russian sign language is focused on the use of UNITY 3D and Dimskis notation. The structure of the system is shown in Figure 1. It allows the developer creating “primitives” of clips that correspond to Dimskis notations, associate them with pictograms of notational signs. The directory of dactyls and sign language words is created on the basis of the obtained library.

The reference book is used for developing lessons and tests. The testing mode includes two types of control:

- The avatar displays a phrase or word, and the user must define them;
- A fingerprint or a word is required, and the system determines the correctness of the gesture shown by the user.

The second test mode is based on gesture recognition. The conventional web camera built into a laptop (Smartphone) is often used. In this case, the environment is static with a sufficient degree of illumination for fixing the gesture with a webcam.

Great attention is paid to the recognition of hand gestures in the framework of man-machine interaction. Gesture recognition technologies include sign language recognition [1], virtual reality applications [2], human-machine interaction [3] and other directions [4, 5]. Most of these areas focus
on the position of the arm and hand, and the set of possible finger configurations is limited, which is extremely important for recognizing sign language. In addition, when testing sign language, the hand appears in the frame for a few seconds, that proposes the preparation of models for comparison in advance. In accordance with the generalized model of information processing within the framework of human-machine interaction [6], four stages of information processing are distinguished: data collection (image capture), data analysis (gesture localization), decision making (gesture identification) and response (output) (figure 2).

![Figure 1. The structure of the integrated sign language teaching system](image)

![Figure 2. Information processing model](image)

Image capture is performed using a motion detector, based on the analysis of binary images of neighbouring frames 32 * 32 pixels in size. With its help, the initial, static with a gesture and the final frames are fixed. The gesture localization is based on subtracting the background (initial frame) from the static and final frames. Then the image in the final frame is blurred, and this frame is subtracted from the static one. All deleted frame points are saved in the alpha channel. Thus, the remaining points of the original static frame should be directly an image representing the sphere of interest. It is put in a
rectangle, and the frame is reduced to its size. Similar actions are performed with the demonstration of a gesture by an avatar, which allows getting the standard image for comparison [7]. The frame received from the web camera is 640 * 480 pixels. The image of a hand with this resolution corresponds to less than 200 pixels vertically, which allows the further reduction of the investigated area size and at the same time to remove the uninformative part of the hand, which may be the significant part of the user depending on his clothes. For horizontal gestures, the image is cut off horizontally.

In contrast with the classical concept of classification (assigning of the analyzed gesture to one of the previously known classes), the system needs to compare (identify) the received gesture with one predetermined gesture. Many classical classification methods [4] use preliminary training, which is not required in the developed system. The most SL gestures are dynamic ones. The recognition of such gestures is based on tracking of the movement trajectory between static gestures. Thus, the primary problem is the recognition of a static gesture.

2. Gesture recognition based on comparison with the standard
The qualitative identification of the gesture is hindered with a number of difficulties. These are the remnants of the background, random and local interference, illumination and backlighting of the gesture, the difference in size and orientation of the standard and the identified gesture. For this reason, the image obtained at the localization stage differs from the standard in geometric and brightness distortions, as well as the remaining noise.

The identification methods can be divided into three groups. The first group is based on direct comparison with the standard, the second one is based on the selection of features and processing in this space, and the third one - on the study of the "design" of the images in question (syntactic recognition). The first two groups of methods are of then combined into one group.

For solving recognition problems, three approaches are mainly used:
Correlation. The approach based on decision-making depending on the criterion of proximity with standards. It is a laborious approach from the point of computing resource consumption.

Indicative. Such methods are based on the transition to the trait space and require significantly less computing power. Then the correlation processing of features obtained from the reference and the input image is performed.

The syntactic method is based on the obtaining of structural and grammatical features when non-derivative elements - features are highlighted in the image. The rules for connecting these elements, which are the same for the standard and the input image, are introduced. An analysis of the grammar obtained in this way ensures decision making.

Characteristic and syntactic methods are most often used in identifying gestures.

In connection with the restrictions listed above, the attribute method is used in the system under consideration, as it requires less resource. The complexity of its application is the selection of features. The signs should be able to identify the gesture and, at the same time, their number should be small. The comparison procedure consists in calculating the inter-correlation function of the gesture with the standard. The sensitivity threshold, determining the minimum value of the similarity function, which decides the correspondence of the identified gesture and the standard, is determined empirically. An additional step which makes the final decision on the identification of the gesture may be required.

The most common and quite effective are the projection recognition method and the recognition method based on the analysis of two-dimensional histograms.

The determination of the maximum of the similarity function can be carried out according to the projections of bitmap gestures and standard. If the similarity between the unknown object and the standard is large enough, then this object is marked as corresponding to the reference object. The full coincidence of the standard with the image is rarely due to the effects of noise and distortion. The method has high recognition accuracy for a given set of standards, even in the presence of random noise. The important advantage of the method is that the standards are set directly in the form of bitmaps; therefore, the additional costs for the standards preparation are not required. The main
disadvantage of the comparison method with the standard is the need to use a certain number of standards to account for changes in objects that occur when they are rotated. In addition, it is necessary to put the standard and the identified object to the same size.

The histograms represent the distribution of image pixels by brightness (by levels), that is, they describe the frequency of occurrence of individual elements values (pixels) independently of others. The advantage of the method is that the frequency values do not depend on the spatial distribution of image elements, that is, the appearance of the histogram does not change upon rotation or shift of the object. Histograms are also often used to localize a gesture.

For carry out the identification of a gesture, four metrics are most often used as a measure of comparison between the image and standard. They are the Pearson correlation coefficient, chi-square, intersection, Bhattacharya distance.

3. Gesture identification in test mode

The localized image is cut off vertically or horizontally depending on the gesture up to 200 pixels. 2 projections on the X and Y axis are calculated. After this, the additional removal of the remaining background and noise is carried out. An example of processing the fingerprint of the letter “H” is shown in Figure 3.

![Figure 3](image)

**Figure 3.** The localized user’s hand and the horizontal projection of the “H” fingerprint before removing extra pixels and after deleting

Before identification the scaling is carried out, the projection of images is reduced to the same size. The direct gesture identification, comparison with the standard may take place in several stages, depending on the intermediate results.

The first stage of identification is the comparison of the projections of the hand. The example of the comparison of the fingerprint "Ж" is presented in Figure 4.

![Figure 4](image)

**Figure 4.** Compared images of the fingerprint "Ж" and their projection

The correlation coefficient was calculated for analysis. Horizontal: $d_{\text{hor}} = 0.558517943$ and vertical: $d_{\text{ver}} = 0.525141243$.

The second stage is a comparison of only the fingers and only the hand vertically. If we perform vertical calculations in parts for the fingers and the hand (the projection was divided in half), then the correlation coefficient is respectively: only fingers – $d_{\text{ver1}} = 0.981437897$, only hand – $d_{\text{ver2}} = 0.173005802$. The second stage is compulsory and supplements the information on the first stage. All analyzes were performed in all four metrics in order to assess the accuracy and implementation time. The results showed that the accuracy of the determination is sufficient for all metrics, and the
calculation time is almost the same. This result is because the calculations of intersection and chi-square coefficient require the additional projections normalization.

Correct gesture identification often requires the third stage – a separate analysis of fingers configuration. This process is due to the fact that the fingers configuration plays the main role in gesture identification. The image is cut off in half vertically, while the configuration of the fingers remains (Fig 5).

According to the data obtained from the three stages, the final decision on the coincidence of the gesture shown by the user with the standard (avatar) is made. The adoption of coincidence decision is carried out in the following sequence:

1. If the vertical correlation is weak (first and second stages) - a message about poor recording is displayed.
2. If it was possible to identify coincidence in the first two stages, a message indicating the correct answer is displayed.
3. If the first two stages did not give a clear answer, then in the third stage, it is specified whether it is the correct answer or not.
4. With the correct answer, the hand histogram is constructed for gestures filtration against the background of a person.

4. Features of the test mode
The difficulty of gesture highlighting against a person's background is that he himself is not a static background, the user's clothes can be quite contrasted, and when a gesture is demonstrated, the shadow falls on it, which changes the colour of the clothes. Therefore, when highlighting a gesture against the user's background, the histogram-based highlighting is additionally used. Figure 6 shows an example of hand selection.

![Figure 5. Fingers configurations projections of user's and avatar fingerprint "Ж"

3. If the first two stages did not give a clear answer, then in the third stage, it is specified whether it is the correct answer or not.
4. With the correct answer, the hand histogram is constructed for gestures filtration against the background of a person.

4. Features of the test mode
The difficulty of gesture highlighting against a person's background is that he himself is not a static background, the user's clothes can be quite contrasted, and when a gesture is demonstrated, the shadow falls on it, which changes the colour of the clothes. Therefore, when highlighting a gesture against the user's background, the histogram-based highlighting is additionally used. Figure 6 shows an example of hand selection.

![Figure 6. Highlighting the hand against the background of the user: a – selected object under investigation; b – use of a histogram for hand highlighting; c – cut off image horizontally up to 200 pixels](image)

In some cases, additional actions are required in the form of the standard rotation. This fact is because when demonstrating a gesture, the user's hand may not be strictly vertical or horizontal. This changes the image projection greatly. An example of the dactyl "Ы" demonstrates this fact (Figure 7).
5. Conclusion
The proposed approach to the analysis of the gesture both the avatar and the user is universal. It allows test any newly developed gesture in the system. For speed up the identification of the user's gesture, it is more rational to prepare the avatar gestures in advance and store them in the database. It is also necessary for some gestures to store rotated standard gestures. The measures considered for comparing the standard and user gestures have the same identification time. To decide on a poor-quality image of a gesture (very noisy, blurred), the first two stages of identification are sufficient. For the correct decision on the coincidence of gestures, it is necessary to perform all three stages of identification.

Acknowledgment. The reported study was funded by RFBR and DST according to the research project No. 19-57-45006.

References
[1] Burger T., Urankar A., Aran O. Akarun L., Caplier A. 2007 Cued speech hand shape recognition — belief functions as a formalismo fuse svms and expert systems VISAPP 2007: 2nd Conference (International) on Computer Vision Theory and Applications Proceeding INSTICC Press., Vol. 2 pp 5–12.
[2] Aguiar R., Pereira J. M., Braz J. 2009 Gadevi — game development integrating tracking and visualization devices into virtools GRAPP 2009: 4th Conference (International) on Computer Graphics Theory and Applications Proceedings INSTICC Press pp 313–321.
[3] Jacob, M., Li, Y.-T., Akingba, G., Wachs, J.P. 2011 Gestonurse: a robotic surgical nurse for handling surgical instruments in the operating room Journal of Robotic Surgery 6 53–63.
[4] Hongyi Liu, Lihui Wang 2018 Gesture recognition for human-robot collaboration: A review International Journal of Industrial Ergonomics Elsevier 68 355-367
[5] Hartmann, F., Schlaefer, A. 2012 Feasibility of touch-less control of operating room lights International Journal of Computer Assisted Radiology and Surgery 8 pp 259–268.
[6] Parasuraman R., Sheridan T. B. Wickens C. D. 2000 A model for types and levels of human interaction with automation Systems, Man and Cybernetics, Part A: Systems and Humans IEEE Transactions on 30 286–297.
[7] Grif M. G. Lukoyanychev A. V. 2019 Gesture localization in the test mode in the integral system of sign language training/ Journal of Physics: Conference Series, Volume 1333, 032023.