Gesture localization in the test mode in the integral system of sign language training

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 article is devoted to the integral system of sign language training. The operation modes of the system based on 3D computer avatar are considered. The user’s feedback using the system is carried out in the test mode using a standard webcam. The user must demonstrate the word or dactyl specified by the system. The problem of gesture localization in the user test mode is considered. The algorithm for the rapid clustering of human hand gestures in a video sequence is proposed. The background removal is based on finding of the inter-element difference for the frames of a video sequence, the selection of the key frames and calculating difference frames. The use of the modified histogram method allows localizing the gesture. The experimental results show the stable operation of this algorithm in different light conditions. This approach allows using developed system both on stationary computers and on mobile devices.

1. Introduction
Multimedia applications improve the visibility and increase the efficiency of human-computer interaction. They can be successfully used in teaching deaf and hearing-impaired people to sign language (SL) instead of video films, sign language interpreters and books. Promising systems in this direction are integral systems that allow not only demonstrating SL gestures, but also having feedback that allows you to check the adoption of the material. The proposed integral learning system of SL contains an editor for creating demonstration animations with dactyl, words and phrases of sign language, a viewer for the created material review. This article considers the creation of the user testing subsystem on the developed topics.

2. Operation modes of the sign language training system
The use of 3D computer character (avatar) is the possible option of SL for computer synthesis. To implement the 3D sign language interpreter, a powerful Unity3D graphics engine was chosen, cross-platform and accessibility being its main advantages.

Several notation systems are used to represent a sign language in a form suitable for computer entry [1]. They have different degrees of detail and iconicity and are focused on the description of various phenomena in sign languages. The notation developed by L. Dimskis, which uses the “photographic” symbols of the fingers configuration, is promising for the Russian sign language [2].

SL training system with two modes: “Developer” and “User” has been developed. In the “Developer” mode the technological tool-kit includes all the necessary modules for gestures creating and
demonstrating, the connection of developed clips to the Dimskis notation, creating the ultimate multimedia SL reference book, and the development lessons for teaching Russian SL.

In the "User" mode the modules for the dactyl library and words review, as well as the created lessons are available for downloading to personal computers and mobile devices. Figure 1 shows the available modules of the software system and the example of the viewer appearance available to the user.

The first two modules “Creating a Primitive Clip” and “Linking of Clip and Images to Notation” are available in the “Developer” mode only on the local computer. The following two modules: “Creating signs based on notations” and “Tests” (creating lessons) are also available to third-party developers after authorization (registration) on the site. The modules "Words view" and "Lessons" can be downloaded to the user device or available in the Internet.

Figure 1. The appearance of the software system and the viewer of dactyl and words directory

The developed integral training system has two modes: demonstration with dactyl, words and phrases of SL (reference book and lessons) and knowledge testing (test mode). The test mode has two test options. The first option: the avatar demonstrates a gesture (word, dactyl) and the user must type the answer text on the keyboard. The second option: the avatar demonstrates a gesture or the system asks to show a gesture corresponding to the dactyl or word, and the user must either repeat it or show it himself. The second test version is based on the recognition of the user’s gesture.

3. Gesture localization methods

Hand gestures recognition is one of the central tasks in human-machine interaction (Human-Robot Collaboration). There is a large number of approaches for solving this problem. These approaches can be classified according to the type of input data: those using visual information or specialized non-visual sensors [3].

The methods focused on visual information are also divided into several directions: the placement of any markers on your hand, the use of deep-sensing, stereo or ordinary cameras.

There are no special requirements to the developed system. For this reason, the use of additional equipment to improve and simplify the gestures recognition, such as special gloves, color markers or deep-sensing cameras, is not provided. The focus is on the usual, often built-in webcam. Modern webcams have significantly improved quality indicators and are sufficient suitable for this purpose. The main disadvantages of this approach are the substantial dependence of the image quality on the illumination and the limitation of the viewing angle. The webcam is able to adapt to the light within certain limits, which allows to localize (detect) the hand, and the user is located directly in front of the camera and sees the image of his hand.

Gesture recognition consists of two stages: the detection of a hand in the image and its localization (search for the area of interest) and the direct recognition and classification of a gesture (the comparison with reference images on the selected characteristic gesture features or other parameters describing the configuration of the hand). When recognizing gestures, it is necessary to detect the human hands on the image. If the gesture is dynamic, it is necessary to record the change in the position of the hands over
time. For this purpose, the approaches based on the detection of areas with human skin color [4, 5, 6], hand movement detection [7], or both are used together [8, 9].

There are no special requirements for the environment and lighting. For this reason, it is difficult to use the method of gesture selection by skin color, without complicating the system. The test question takes a few seconds to complete. During this period of time, the illumination and the environment can be considered invariable. For this reason, the most effective method of image pre-processing is background removal. During this time there also will be no moving objects in the frame, except for the hand. Thus, the motion detection method is used to detect a gesture.

The peculiarities of the work of SL training system in the test mode include the following:

- In the test of knowledge, firstly, the avatar shows a gesture, and the user must repeat it or, secondly, the user shows the gesture, and the avatar shows the correct one. Thus, the answer is known and the stage of the gesture classification is reduced to its identification.
- The user, when testing knowledge, is in a calm atmosphere in front of the webcam and we can consider that there is a static background behind it. However, in this case there is no restriction on lighting and surroundings, which complicates the recognition task.
- Before to start the hand movement, the user must turn on the “Test Display” mode, that is, to press the video capture button. At the end of the show, the user must turn off this mode (timer-defined mode is possible). Thus, the process of image processing has a clear start and end.
- Co-articulation (continuous articulation, merging of gesture phases) is a characteristic feature of sign language. The approach based on the selection of palm position data is used to recognize dynamic gestures and determine co-articulations. Even with a dynamic gesture, the palm with the certain configuration of fingers should remain in a static state for a fraction of seconds.
- The main, ideally the only moving object is the hand, which makes it easier to detect a gesture. The hand is in the foreground of the webcam, with a slight change in the illumination of remote objects. Besides, the user's body and head also move within certain limits. This does not allow to consider the body and head of a person as a background. A clear knowledge of the end of the test allows you to use the last frame to re-correct the removal of the person image.

As a result of such actions, only one object of interest remains on the image - the hand. The object localization means determining the presence of an object in the image, defining its position coordinates and highlighting its location area. The depending on the localization algorithm, the position of the object can be determined in different ways.

Localization of the object of interest can be made on the basis of the form, the key characteristic features that significantly distinguish the object from the bulk of points, or on the basis of color and brightness characteristics. To determine the gesture of a hand which has no specific form and special features, it is possible to use only the latter approach.

The histogram method may be used to detect and localize the hand in a fast way that does not require a lot of computations [10].

To detect an object, histograms are built according to the criterion of color or brightness uniformity, and in order to localize the object histograms are built according to the number of bright dots in the image. In both cases the background must be previously removed. The basis of the method is the selection of the range of these parameters for certain threshold values during image processing, which allows you to select areas of a certain type. The separation boundary is defined in two ways: with a global threshold and with an adaptive threshold.

The simplest of the threshold processing methods is to divide the image histogram into two parts using a single global threshold. If the detection is done by brightness, then the image is usually converted into a halftone image. The image is then scanned element-by-element, with each pixel marked as referring to the object (hand) or the rest of the background, depending on whether the brightness of given pixel exceeds the threshold value or not. The success of this method depends entirely on how well the image histogram can be divided.
The histogram method is also used to localize the object under study. To determine the area boundaries of the object of interest, the histogram method is to build two image histograms. Histograms are plotted along the X and Y axes by the number of bright dots. As a result of the double application of the method, areas not related to the object of interest are cut off.

The method depends on the lighting, filtration methods and, therefore, the processing system must be highly adaptable to changes in the external environment. The popularity of methods based on histograms is explained by the simplicity of calculations, resistance to movements, rotations of the object of interest and changes in the camera position. The use of this method precedes the process of direct image analysis and identification.

4. Gesture localization in the test mode
When you switch to the test mode, the webcam is turning on. For example, in the test mode, the system asks the user to repeat the gesture shown by the avatar. From this moment (or by pressing the key) the video capture of the image begins. The system captures the first frame that is used to register the background. After that the motion detector turns on.

The calculation of the interframe difference for the case of color video processing in RGB format consists in taking of two consecutive frames from a webcam to the input. To speed up processing, the original full-color RGB image is reduced to the binary image. Initially, the image is converted to a halftone image with brightness I using the standard transformation:

\[ I = 0.2125 R + 0.7154 G + 0.0721 B \]

where \( R, G, B \) are red, green and blue pixel values, respectively.

Then the image is transferred to the binary image in accordance with the threshold filtration value. The threshold value can be either a constant or a variable depending on the total illumination. At low lighting the threshold lowers, at the light background it increases that allows detecting movement more reliably. Frame sizes are compressed to a resolution of 32 x 32 pixels. Two adjacent binary frames are compared. Per-pixel interframe differences are calculated according to the following scheme:

\[ D^*(x, y) = D(x, y, n-1) - D(x, y, n) \]

where \( D(x, y, n) \) are the brightness values for the current pixel of the n-th frame of the video sequence; \( D(x, y, n-1) \) - brightness values for the current pixel of the previous frame of the video sequence.

\[ S = \sum_{x, y} D^*(x, y) \]

is calculated. The condition \( S > T \) is analyzed, where \( T \) is the filtration threshold. If the frames differ by more than \( T \), then the beginning of movement is recorded. The detector switches to the end-of-motion fixing mode. If the adjacent frames do not differ by more than \( T \) within 0.25 seconds, the user shows a static gesture. A fixed static frame is recorded. The detector comes to the initial state to determine the new movement, that is, it switches to fixing the next static frame. With a dynamic gesture there are several such frames. The demonstration of a gesture takes 4-5 seconds, or the end of the gesture is fixed by pressing the button. On the completion of the gesture demonstration, the final static frame is recorded. It is used to eliminate the background remnants. This is due to the fact that the user may move during the demonstration, and therefore the first and the last frames will be slightly different.

At the end of the recording, the gesture localization stage begins. The first frame (background) is subtracted from all the recorded frames (Figure 2).

The last frame is subtracted from the intermediate frames (the dactyl “B” is shown in the example), but previously it was subjected to the blur operation in order to eliminate reliably the background residues (figure 3).

The change in the background is due to the presence of unpredictable movements of the human figure.
As a result, the palm, hand and remnants of the background remain. The largest area with the greatest brightness is occupied by the hand and, possibly, part of the arm, depending on the user's clothes. The histogram method is then used to highlight the area of interest. Two histograms are constructed horizontally and vertically. The image is divided into N parts along the X axis and Y axis, the number of bright dots is calculated in each part. Splitting allows you to exclude the analysis of the rupture of the hand image contour the localization stage. The areas where such points are less than Z, where Z is a threshold value, are cut off. As a result, the object of interest is limited to a rectangle (figure 4a).

The same actions are performed for the avatar (figure 4b). The use of the implementation capabilities in Unity3D allows you to select only the palm with the fingers, which defines a significant part of the gesture. As a result, only the hand remains for the avatar in the selected area (figure 4c). At this point the localization stage finishes. And further actions are associated with the next stage - the identification of the gesture, the comparison of the localized image of the user's hand with a similar avatar image.

5. Conclusion
The proposed algorithm of gesture localization works steadily in low light, both of the environment (background) and of the user's workplace, and allows two-handed gestures to be selected (figure 5). The conversion of image to grayscale in the early stages of localization allows speeding up image processing. The other way of acceleration is the reduction of the size of the processed frames. The experiments have shown that 340 * 256 pixels are sufficient for stable localization. The use of double removal – of the background and the final frame, the use of a quick method of the area of interest localization based on histograms allows you to apply the proposed algorithm, both on stationary devices and on mobile ones.

Figure 5. Localized images at low (a) and medium (b) illumination of the user's workplace, at low illumination of the background environment (c) and localization of the two-handed gesture (number seven) (d)

References
[1] Grif M G, Prihodko A L 2018 Approach to the Sign language gesture recognition framework based on HamNoSys analysis Proc. of the 14th International Scientific and Practical Conf. Actual problems of electronic instrument engineering 2–6 Oct. Novosibirsk vol 1 part 4 (Novosibirsk: NSTU) pp 426-429
[2] Prihodko A L, Lukoyanov A V, Grif M G 2016 Approach to the Analysis and Synthesis of the Sign Language Proc.13 International Scientific. Technical Conf. on Actual Problems of Electronic Instrument Engineering 3–6 Oct. Novosibirsk vol 1 part 2 (Novosibirsk: NSTU) pp 502-505
[3] Hongyi L, Lihui W 2018 Gesture recognition for human-robot collaboration: A review International Journal of Industrial Ergonomics 68 355-367
[4] Phung S L, Bouzerdoum A, Chai D 2005 Skin Segmentation Using Color Pixel Classification: Analysis and Comparison IEEE Trans. Pattern Anal. Mach. Intell. 27 148–154
[5] Vezhnevets V, Sazonov V, Andreeva A 2003 A survey on pixel-based skin color detection techniques. Proc. Graphicon-2003, September pp. 85–92.
[6] Bretzner L, Laptev I, Lindeberg T 2002 Hand Gesture Recognition using Multi-Scale Color Features, Hierarchical Models and Particle Filtering Proc. of the 5th Int. Conf. on Automatic Face and Gesture Recognition
[7] Cutler R, Turk M 1998 View-based Interpretation of Real-time Optical Flow for Gesture Recognition IEEE Int. Conf. on Automatic Face and Gesture Recognition. – Nara (Japan)
[8] McKenna S, Gong S 1998 Gesture Recognition for Visually Mediated Interaction using Probabilistic Event Trajectories Proc. of BMVC’98. – Southampton (England)
[9] Marcel S, Bernier O, Viallet J-E, Collobert D 2000 Hand Gesture Recognition Using Input-Output Hidden Markov Models Proc. of the Fourth IEEE Int. Conf. on Automatic Face and Gesture Recognition
[10] Otsu N 1979 A Threshold Selection Method from Gray-Level Histograms IEEE Transactions on Systems, Man, and Cybernetics 9(1) 62-66