Dynamic Hand Gesture Tracking Method Based on Key Frame and Video Perceptual Hashing

Qiuyu Zhang, Qian Wang, Moyi Zhang and Hongxiang Duan
School of Computer and Communication, Lanzhou University of Technology, Lanzhou, 730050, China
Email: zhangqylz@163.com; 913680614@qq.com

Abstract. Aiming at the common problems such as low tracking accuracy, robustness and poor real-time ability for the traditional tracking algorithm, a novel dynamic hand gesture tracking method based on key frame and video perceptual hashing has been proposed in this paper. Firstly, the proposed method is used to frame the video sequence and extract the key frames, in order to obtain a sequence of key frames. Secondly, the video perceptual hashing is used to extract the perceptual feature of the key frame and target template, thus the hash sequence is generated. Finally, the hamming distance is used to compare the similarities between the key frame and target template, therefore, the real-time tracking is carried out by using an online adaptive template update method to ensure the continuity of the hand gesture tracking. Experimental results show that the proposed method can realize the accurate and real-time tracking of the dynamic deformation gestures, and effectively reduce the time complexity of the algorithm, as well as ensure the stability and continuity of the moving target tracking.

1. Introduction
With the development of human-computer interaction technology, the dynamic hand gesture tracking problems has become one of the core contents of hand gesture interaction [1].

 Currently, the commonly used methods of hand gesture tracking mainly include the Mean-shift, Cam-shift and other tracking methods, with the target occlusion, illumination changes and color interference the hand gesture target by using these methods easy to lose and the tracking is not accurate [2]. However, the introduction of perceptual hashing technology in target tracking provides a new idea and method for the research of hand gesture tracking [3-13], which can effectively solve the problems of the above methods, and it has stronger stability and robustness. Yu [3] proposed a tracking method based on Discrete Cosine Transform (DCT) image perceptual hashing algorithms; it compared the hash value of the target with the next frame image, so as to achieve the purpose of object tracking. Li [4] used the perceptual hashing technique in vehicle tracking to extract the feature of the target vehicle, in order to identify and match the hash value of the target. Li [5] obtained the image hash value through the DCT technique, and compared the similarity between the target model and the current frame, so as to track in real-time. Fan [6] extracted the perceptual features of each frame images, and generated the hashing fingerprint, as well as tracked the target by comparing the hashing fingerprint of each frame images, in order to track the object effectively. Shen [7] used a kind of online perceptual hashing and color self-similarity block feature frame in the process of feature extraction. Fei [8] put forward a Laplace-based Hash (LHash) and Laplace-based Difference Hash (LDHash) on the basis of traditional perceptual hashing technology, these two methods can track the target effectively in real-time. They also used the average hash (aHash), perceptive hash (pHash) and difference hash (dHash) techniques in visual tracking, and compared with the traditional tracking
methods with the LHash and LDHash, finally the experimental results have indicated that the perceptual hashing in visual tracking has stronger robustness and real-time performance. Sengar [9] tracked the motion object by using the Laplace-based Difference Hash (LDHash), the method can track the target effectively with scale transformation. Liang [10] computed the hamming distance between the current frame’s hashing fingerprint and the last frame’s hashing fingerprint, in order to track the target by using the similarity of two frame images. Zhao [11] used the hashing distance between images in the template updating strategy, and judged whether to update the template or not determined by the similarity of images, so as to match and track the next frame images with current frame images. Xu [12] used the perceptual hashing to extract the perception features of each frame images and generated a hashing fingerprint, so as to track the target effectively. All above researchers extracted the perceptual features of each frame images by using perceptual hashing technology, this tracking method had stronger robustness, and also satisfied the real-time performance. But the method required extracting the perceptual features of each frame images in the video sequence, it led to the unnecessary perception features extraction for moving objects which has not large changes in the video, and increased the computation of the method.

To solve the above problems, this paper proposes a new dynamic hand gesture tracking method which is based on key frame and video perceptual hashing. Firstly, the video is framed and key frame are extracted from the video sequence. Then, taking the perception features of key frame by using video perceptual hashing technology, and calculating the hash value of key frame. Finally, comparing the similarity between key frames and the target template by using the hamming distance, so as to track the hand gesture target in real-time. As well as an online adaptive template updating method is used to ensure the continuous tracking of the dynamic hand gesture in complex environment, and experiment results show that the proposed method can reduce the computation and satisfy the real-time requirement.

2. The Proposed Method
The flow chart of dynamic hand gesture tracking method based on key frame and video perceptual hashing are shown in Figure 1.

![Figure 1. Flow chart of the proposed method.](image)

2.1. Extracting Key Frames
Step 1: Calculating Euclidean distance between frames as difference degree between adjacent frames. The formula is:

$$
\text{diff}_n = \sqrt{\sum_i \sum_j (G_n(i,j) - G_{n-1}(i,j))^2}
$$

(1)

where $G_n(i,j)$ is defined as:

$$
G_n(i,j) = R_i \times R_n(i,j) + G_1 \times g_n(i,j) + B_i \times b_n(i,j)
$$

(2)
where $R_1 = 0.299$, $G_1 = 0.587$, $B_1 = 0.114$, $r_n(i, j)$, $g_n(i, j)$ and $b_n(i, j)$ respect the red, green, and blue components of $n$th frame image in $(i, j)$.

**Step 2:** Selecting the maximum values of $\text{diff}_n$ ($n = 2, 3, \ldots, N-1$) as the key frame sequence, and the judgment formula is:

$$
e(i) = \begin{cases} 1 & \text{if } \text{diff}_n(i) \cap \text{diff}'_n(i) = 0 \\ 0 & \text{otherwise} \end{cases}$$

(3)

**Step 3:** Sorting all of the maximum points to select the $M$ biggest points as key frames, that is to say, these key frames between frame images has changed greatly. Setting $M=30$.

**Step 4:** Ensuring the order of the selected $M$ frames by using time-warped algorithm so that the final key frame of the video image sequence is strictly chronological.

### 2.2. Generating Hash Value for Target Template

Using the discrete cosine transform (DCT) to the target template, and generating a hash value sequence by extracting the DCT coefficients and quantum hashing, as well as the main steps are as follows:

**Step 1:** Taking interpolation processing to the target template, and resizing the image to $32 \times 32$, in order to reduce the computation of the latter DCT transform.

**Step 2:** Converting the image to the grayscale level, and further reducing the computation.

**Step 3:** Taking the DCT transform for the image and obtaining a $32 \times 32$ DCT coefficient matrix.

**Step 4:** Retaining the upper left corner $8 \times 8$ matrix of the $32 \times 32$ DCT coefficient matrix, and calculating the mean value of the DCT coefficient, then using the coefficient to quantize the matrix. And setting the coefficient which is greater than or equal to average as “1”, less than the average coefficient is set to “0”.

**Step 5:** Reading the quantum coefficient matrix line by line, and generating a length of 64 one-dimensional vector $A_i$. And $A_i$ is the hash value for the target template, which is composed of “0” and “1”.

### 2.3. Hash Matching

The hash value is a binary string with a length of 64 bits, namely, it is the hash fingerprint which can represent the information for the original image, and judge the similarities between the two images by matching the fingerprint. The hamming distance is usually used to match the hash value and calculate the distance between the template and the hand gesture target. And the formula is:

$$\text{Sim} = \text{HamDis}(h, h')$$

(4)

where $h$ and $h'$ respectively represent the hash value for the template and key frames. And $\text{HamDis}$ is the hamming distance calculation function, $\text{Sim}$ is the matching similarity.

In matching the hash value, it need to set a threshold $T_{\text{Sim}}$, when $\text{Sim} \leq T_{\text{Sim}}$, it means that the matching is considered to success, that is to say, the two images are very similar. And when $\text{Sim} > T_{\text{Sim}}$, it means that the matching is considered to fail, namely, the two images has widely difference.

In the process of matching the key frame’s hash value, because of the large difference between adjacent key frames, the threshold is set to $T_{\text{Sim}} = 10$ after several experiments and verification, namely, it can satisfy the requirements of the experiment.

### 2.4. Adaptive Updating for Target Template

This paper adopt an adaptive template update method, when the key frame is obtained and get ready to track it, the first frame in the video is selected as the starting template and matched with the first key frame. After it has finished matching, the first key frame is selected as the next template, and matched with the next key frame. Until all the key frame has matched, it need to match the last key frame with the last frame in the video.
When $\text{Sim}(FK_1) \leq T_{\text{Sim}}$, it is considered that the $1^{\text{st}}$ frame in the video is similar to the $1^{\text{st}}$ key frame, and then the $1^{\text{st}}$ key frame is selected as the template and matched with the next key frame. On the contrary, the matching is failed. Where $F$ represent the $1^{\text{st}}$ frame in the video, $K$ represent the key frame and $K_1$ represent the $1^{\text{st}}$ key frame.

When $\text{Sim}(FK_n, K_{(n+1)}) \leq T_{\text{Sim}}$ ($n \leq 29$), It is considered that the $n^{\text{th}}$ key frame is similar to the $(n+1)^{\text{th}}$ key frame, and then the $(n+1)^{\text{th}}$ key frame is selected as the template and matched with the next key frame. On the contrary, the matching is failed.

When $\text{Sim}(LK_{30}) \leq T_{\text{Sim}}$, It is considered that the last frame in the video is similar to the last key frame, that is to say, the matching is succeed. On the contrary, the matching is failed. Where $L$ represent the last frame in the video.

3. Experimental Results and Analysis

In order to verify the validity of the method in this paper, using OpenCV 2.4.9 to achieve the tracking method of Ref. [5] in VS2013 environment, and finishing the traditional Mean-shift algorithm in Matlab2013a, and also implementing the proposed method in the Matlab2013a and OpenCV 2.4.9. And compared the difference between the three methods’ performance in testing. The PC’s configuration is: Intel Core(TM) i5- 4590M CPU 3.30 GHz and 4GB. This method use two video sequences (Video 1 and Video 2) to contrast and verify the robustness and real-time of the experiment, the main purpose is to verify the situation of dynamic gesture tracking in different conditions, and under the premise of ensuring the robustness of the video perceptual hashing, in order to research the real-time about the proposed method. And in the process of hand gesture tracking, the first frame and the last frame in the video are added to the sequence of key frames, in order to ensure the stability and invariability of the video content.

3.1. The Extraction of Key Frame

The method of inter-frame difference degree are used to obtain the key frame sequence from Video 1 and Video 2, which can represent the video content, and it is shown in Figure 2.

![Figure 2. Key frame extraction based on different degree value.](image)

3.2. Comparison of Tracking Results

Video 1 (640×480 pixels, 30 frames/s) is a gesture of video sequence in simple background, the feature distinction between the gesture and the background is more obvious, and the gesture is to take simple deformation movement. Figure 3 is the comparison of different tracking methods which is the method of Ref. [5], the Mean-shift algorithm and the proposed method.

As can be seen from Figure 3(a), the method of Ref. [5] has stronger robustness in the experiment, the tracking is accurate, and it is not affected by the changes of the hand gesture. In Figure 3(b), the traditional Mean-shift tracking method can also track the hand gesture target, but the robustness is poor, and the red rectangular box cannot track the whole hand gesture target. And part of the hand...
gesture area is outside the red rectangular box, furthermore, the tracking accuracy is lower. However, the proposed method has nice robustness in the experiment, and in Figure 3(c), in the case of the deformation of the hand gesture, the proposed method can accurately track the target, and it is also not affected by the changes of the hand gesture, as well as it can steadily and continuously track the dynamic hand gesture.

Figure 3. Comparison of different methods for Video 1.

The traditional Mean-shift algorithm use the method based on color histogram statistics, and find the optimal location of target gesture by iteration. But it cannot accurately track the target gesture, and it is prone to drift track or lost target gesture with illumination changes, skin color interference and the deformation of the hand gesture. The method of Ref. [5] use perceptual hashing to extract the features for the gesture target and the template, match the hash value with the target and the template, and ensure the tracking accuracy by using an adaptive template update method. So it has better resistance, and has stronger robustness. But it needs to compute the hash value for each of frames in the video, and need to match and track all the frames in the video, so it has largely computational complexity. However, the proposed method only need to calculate and match the hash value for key frames, use an adaptive tracking method to update the template, and it has stronger real-time performance under the premise of robustness with the video perceptual hashing.

Video 2 (640×480 pixels, 30 frames/s) is a gesture of video sequence in complex background, and compared with Video 1, the video scene is chaotic and the gesture continuously changing. Figure 4 is the comparison of different tracking methods which is the method of Ref. [5], the Mean-shift algorithm and the proposed method.

In Figure 4(a), the method of Ref. [5] can still accurately track the deformation of the hand gesture with illumination changes, color interference and scale changes for the hand gesture target. It has stronger robustness, but also need to extract and match the hash value for each of frame images in the video. The Mean-shift method is usually affected by illumination changes, color interference and the
deformation of the hand gesture, it can be seen from Figure 4(b), in the 206th frame, the red rectangular box generate the problem of tracking drift due to the changes of the hand gesture. And in the 269th frame, the tracking box begin deviating from the hand gesture target due to the deformation of the hand gesture. Parts of the hand gesture target is tracked by the tracking box, and another parts of the hand gesture target far away from the tracking box. With the tracking continues, the red rectangular box track part of the hand gesture target all the time, causing the inaccurate tracking phenomenon, and it indicate that the Mean-shift method has poor robustness. However, it can be seen from Figure 4(c), the proposed method also can accurately track the deformation of the hand gesture in complex background, and it is immune to skin color interference and illumination changes. So the proposed method also has nice robustness.

(a) The method of Ref. [5]

(b) Mean-shift tracking method
3.3. Real-Time Analysis

In order to compare the time consumption in the process of tracking with the proposed method, the method of Ref. [5] and the Mean-shift algorithm, the time consumption is respectively computed for Video 1 and Video 2, and the comparison of the time consumption is shown in Table 1.

| Tracking method  | The tracking time of Video 1       | The tracking time of Video 2       |
|------------------|-----------------------------------|-----------------------------------|
| Ref. [5]         | 309.42s (343th frame)             | 357.83s (421th frame)             |
| Mean-shift       | 335.23s (343th frame)             | 413.65s (421th frame)             |
| The proposed     | 28.29s (only 32th frame)          | 26.75s (only 32th frame)          |

As can be seen from Table 1, compared with the method of Ref. [5] and the Mean-shift algorithm, the proposed method require fewer time in tracking. And it has lower computational complexity, it can reduce the time consumption for the tracking system.

4. Conclusions

In this paper, a novel dynamic hand gesture tracking method based on key frame and video perceptual hashing is proposed. The experimental results indicate that the proposed method has the advantages of less computation, less time consuming and accurate tracking, comparing with the existing perceptual hashing tracking method, and on the basic of ensuring the robustness of the perceptual hashing technique. The proposed method can quickly and accurately track the hand target, achieve the real-time tracking for the human hand movement, as well as satisfy the robustness. Comparing with the Mean-shift algorithm, the proposed method has stronger robustness and better real-time performance. In addition, this method can also be used to track the target objects.

In the future work, we will study the occlusion and the lost problems of the hand gesture target tracking as much as possible on the basis of ensuring the robustness of the hand gesture tracking.

5. Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 61363078), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (No. 201700005). The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.
6. References

[1] Yeo H S, Lee B G and Lin H 2015 Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware Multimedia Tools and Applications 74, 2687

[2] Feng Z Q, Yang B and Li Y 2012 Hand tracking method based on interactive behavioral analysis Computer Integrated Manufacturing Systems 18 31

[3] Yu S and Jiang Z 2015 Visual tracking via perceptual image hash from a mobile robot Proc. Int. Conf. on Information and Automation IEEE Beijing China pp 1612-1616

[4] Li Z, Yang J F and Chen L 2015 Robust Vehicle Tracking Using Perceptual Hashing Algorithm Proc. Int. Conf. on Machine Learning and Applications IEEE Beijing China pp 1111-1116

[5] Li Z Y, Zhu M L and Chen Z 2015 Object tracking algorithm based on perception hash technology J. of Image and Graphics 20 795

[6] Fan F, Gao G and Li J 2015 Visual object tracking based on perceptual hash algorithm Proc. Int. Conf. on Wavelet Active Media Technology and Information Processing IEEE Beijing China pp 233-236

[7] Shen X W, Sui X and Pan K 2016 Adaptive pedestrian tracking via patch-based features and spatial-temporal similarity measurement Pattern Recognition 53 163

[8] Fei M J, Li J and Shao L 2015 Robust Visual Tracking Based on Improved Perceptual Hashing for Robot Vision Proc. Int. Conf. on Intelligent Robotics and Applications Springer pp 331-340

[9] Sengar S S and Mukhopadhyay S 2016 Moving object tracking using Laplacian-DCT based perceptual hash Proc. Int. Conf. on Wireless Comm., Signal Processing and Networking IEEE Beijing China pp 2345-2349

[10] Liang J P and Zhu X S 2016 Adaptive real-time compressive sensing tracking algorithm J. of Guangxi Academy of Sciences 32 42

[11] Zhao L, Wang Q F and Liu Yuan Y 2016 Target tracking algorithm based on perceptual hash and wavelet transform Systems Engineering & Electronics 38 739

[12] Xu H and He F Y 2016 Hash enhancement algorithm for template matching tracking J. Computer Applications and Software 33 167