TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 15/18 pp329-336
Volume 17, Number 3, June 2012

A Hand Gesture Based Interactive Presentation System Utilizing Heterogeneous Cameras

Bobo Zeng, Guijin Wang**, Xinggang Lin

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Abstract: In this paper, a real-time system that utilizes hand gestures to interactively control the presentation is proposed. The system employs a thermal camera for robust human body segmentation to handle the complex background and varying illumination posed by the projector. A fast and robust hand localization algorithm is proposed, with which the head, torso, and arm are sequentially localized. Hand trajectories are segmented and recognized as gestures for interactions. A dual-step calibration algorithm is utilized to map the interaction regions between the thermal camera and the projected contents by integrating a Web camera. Experiments show that the system has a high recognition rate for hand gestures, and corresponding interactions can be performed correctly.

Key words: hand gesture detection; gesture recognition; thermal camera; interactive presentation

Introduction

Interactive presentation systems use advanced Human Computer Interaction (HCI) techniques to provide a more convenient and user-friendly interface for controlling presentation displays, such as page up/down controls in a slideshow. Compared with traditional mouse and keyboard control, the presentation experience is significantly improved with these techniques. Hand gesture has wide-ranging applications[1]. In this study, we apply it to an interactive presentation system to create an easy-to-understand interaction interface.

The key to this system is robust localization of the presenter's interaction hand. Traditionally, visible light cameras are used for this task. However, in the presentation scenario, complicated and varying illumination by the projector changes the appearance of both the presenter and the hand in the visible light camera, invalidating methods such as detecting hand skin color[2] or training hand detectors[3,4]. To address this shortcoming, we introduce a thermal camera whose imaging is invariant to the projector's illumination.

During the presentation, the interaction hand is small and no salient features exist, so direct localization is not feasible. Hu et al.[5] built a sophisticated human model, but it is burdened by a heavy computational cost. Iwasawa et al.[6] located the head and hand by analyzing the skeleton heuristically. However, only a limited number of postures can be estimated. Juang et al.[7] found significant body points by analyzing convex points on a silhouette contour, but the convex points are not stable when the silhouette's contour is missing. Head localization is important as the first step in many pose estimation methods. Zou et al.[8] extracted quadrant arcs to fit an elliptical head contour, which needs to integrate the arcs combinatorially, so it is still not fast enough.

Calibration of the camera and the projected contents is needed for corresponding interaction regions. Special patterns need to be projected to find the point correspondences in Refs. [9,10], which are inconvenient and not fully automatic. Wang et al.[11] found

Received: 2011-12-30; revised: 2012-05-07

** To whom correspondence should be addressed.
E-mail: wangguijin@tsinghua.edu.cn; Tel: 86-10-62781430

correspondence between shots and slides by utilizing the positions of recognized texts. However, this approach requires the texts to be presented and recognized robustly, which is not easy to do.

1 System Overview

The system setup is illustrated in Fig. 1. It consists of a projector with a projection screen, a thermal camera with 324×256 resolution, and a Web camera. The two cameras are placed facing the screen and close to each other for better correspondence. The presenter stands at the front of the projection screen so the cameras can see the presenter and the screen simultaneously. The presenter uses his/her hand for interaction.

The system flowchart is illustrated in Fig. 2. The thermal camera detects and tracks the presenter's hand (gesture detection), and the hand trajectory is segmented and recognized (gesture recognition). Once predefined gestures are recognized, the corresponding interactions are performed and projected to the screen with the projector-camera calibration. The calibration is computed with the thermal camera and the Web camera together to map the interaction regions.

Fig. 1 System hardware setup

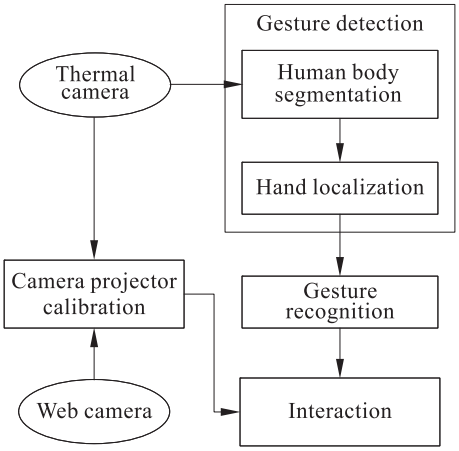


Fig. 2 System flowchart

Robust and fast gesture detection by locating the interaction hand largely determines the system's performance and serves as our major emphasis. Being aware of the difficulties in directly locating hands, we first segment the human body by background modeling with the thermal camera. Based on the segmented human silhouette, we develop a fast hand localization algorithm for our real-time application. The human body is roughly divided into three parts: the head, the torso, and the arm, and they are localized sequentially. The hand is finally localized by skeletonizing the arm.

With the hand trajectory, we segment and recognize three types of gestures (lining, circling, and pointing) and apply the corresponding interactions given the pre-computed calibration. Direct calibration from the thermal camera to the projected contents is impossible, as no projected contents can be viewed from the thermal camera. Therefore, we develop a dual-step automatic calibration method with the aid of a common Web camera. The first step is thermal camera and Web camera calibration based on line detection, and the second step is Web camera and projected contents (e.g., PPT slide) calibration based on Scale Invariant Feature Transform (SIFT) feature points[12].

2 Gesture Detection

We propose a part-based algorithm for localizing the hand point, including a contour-based head detection and shape matching based head tracking module, a histogram-based torso and arm localization module, and a skeleton-based hand localization module. The localization algorithm is based on the observations that the human body is upright and, in interactions, the hand is located away from the torso, either beside the torso (horizontally) or above the head (vertically). The algorithm flow chart is illustrated in Fig. 3.

2.1 Human body segmentation

Human body segmentation is carried out using the thermal camera, as its images are invariant to the projector's light, which is ideal for segmentation. Figure

4c shows a typical image captured by the thermal camera; no content on the projection screen can be observed. We use a Gaussian background model[13] to extract the foreground as the segmented human body. Since the background is simple and static in the indoor environment, unimodal Gaussian distribution is employed.

For post-processing, erosion and dilation operations are performed to remove the noise and connect the possible gaps caused by inaccurate segmentation. Then, we use a connected component analysis algorithm[14] to select the largest connected component as the human body and its contour. Figure 4 illustrates that the human body is well segmented in the thermal camera. The inferior results of the Web camera are also shown for comparison.

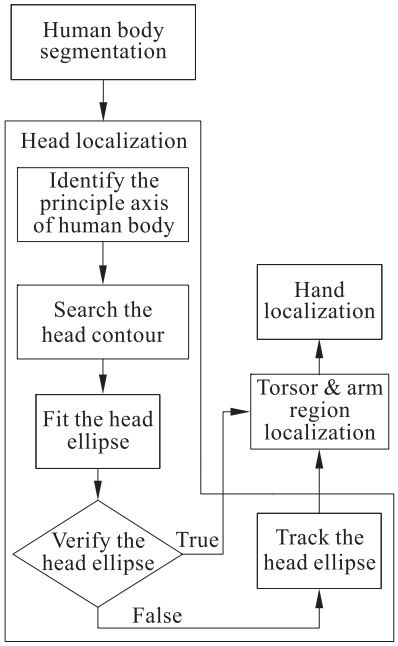


Fig. 3 Hand localization flowchart

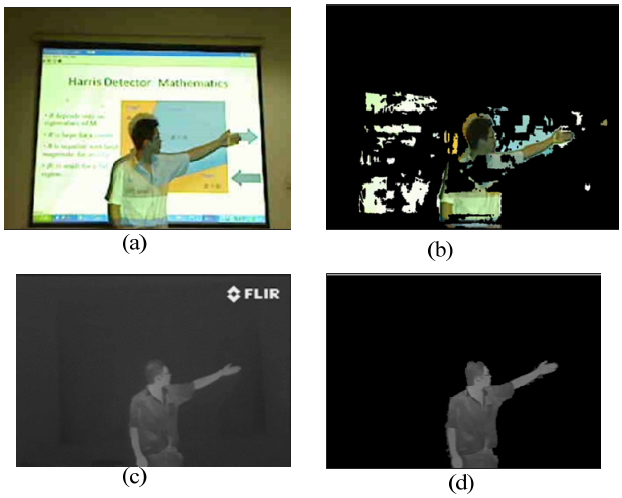


Fig. 4 Comparison of human body segmentation results of the two cameras: (a) image captured by the Web camera; (b) the inferior segmentation result of the Web camera; (c) image captured by the FLIR thermal camera; and (d) the superior segmentation result of the thermal camera

2.2 Head localization

Head localization is performed for three reasons. First, directly localizing the hand is not feasible due to the lack of salient hand features. In contrast, the head has a salient elliptical shape, which is stable and invariant to head poses (e.g., frontal or profile view). Second, the localized head can guide the localization of the torso and arms by its position and size. Third, since the head and arm are quite different in size, the head's localization can prevent it from being falsely identified as the arm. We propose to detect the head by contour-based ellipse fitting and track it by robust shape matching.

2.2.1 Contour-based elliptical head detection

The head contour is segmented from the human body contour by using a gradient angle distribution of its elliptical shape. Then, an ellipse is detected by least squares fitting with the head contour's points. The head contour's gradient angle distribution is illustrated in Fig. 5. Searching from the start point (the highest point on the head contour), the gradient angles of the contour points vary from the first quadrant $[0,\ \pi/2]$ to the fourth quadrant $[-\pi/2,\ 0]$ in the counterclockwise direction, and from the second quadrant $[\pi/2, \pi]$ to the third quadrant $[-\pi, -\pi/2]$ in the clockwise direction. The gradient angle of each contour point is computed by

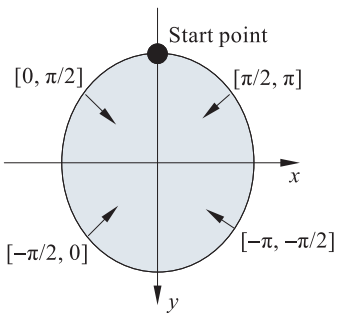


Fig. 5 Illustration of gradient angle distribution of an elliptical head when searching from the start point in clockwise and counterclockwise directions. The arrow line shows the d_x and d_y involved in the gradient angle computation.

$$\theta=\begin{cases}\arctan(d_y/d_x), & d_x>0;\\ \arctan(d_y/d_x)+\pi, & d_x\leqslant 0 \text{ and } d_y\geqslant 0;\\ \arctan(d_y/d_x)-\pi, & d_x\leqslant 0 \text{ and } d_y<0\end{cases} \quad (1)$$

where d_x and d_y are x and y derivatives obtained with a [−1 0 1] mask.

The head detection process is illustrated in Fig. 6. First, the search start point is identified as the intersection of the human body contour and the principal axis of the human body. The principle axis is a vertical line

$$x_p=\arg\max h(x) \quad (2)$$

where $h(x)$ is the horizontal histogram of human body pixels. From the start point, the gradient angles are computed along the contour in counterclockwise and clockwise directions. The image is smoothed by a 5×5 Gaussian kernel before computing the angles to remove noises. The head contour is segmented after traversing the 4 quadrants in order. The symmetry of contour segments in clockwise and counterclockwise directions is employed to reduce false segmentations. If the segment of one side is exceptionally longer than another side, then it is truncated to maintain symmetry.

An ellipse is detected with the segmented head contour points by least squares fitting[15] using the ellipse equation. Then, the ellipse's long axis a, short axis b, and the orientation θ to the y-axis are obtained. The ellipse is verified as a valid head if the following conditions are satisfied:

(a) The fitting error is less than a specified threshold.

(b) According to the head's size, the ellipse's aspect ratio a/b should fall into a range of [1.0, 2.5] and the orientation should fall into a range of $[-\pi/4, \pi/4]$.

If the head is unverified, further head tracking is required. An unverified head is illustrated in Fig. 7, which is caused by a wrongly segmented head contour due to distraction by the arm.

2.2.2 Shape matching based head tracking

When head detection fails, a head template tracking method utilizing robust shape matching is utilized to recover the head. We assume that the head has the same size as the last detected head, but it is now located in a different position. The task is to find the best state hypothesis, $\hat{s}_t=(x_t, y_t)$, which denotes the position of the head at time t. Its probability is measured by

$$\hat{s}_t=\arg\max_{s_t} p(s_t\mid o_t) \quad (3)$$

$$p(s_t\mid o_t)\propto p(o_t\mid s_t)p(s_t\mid \hat{s}_{t-1}) \quad (4)$$

where $p(o_t\mid s_t)$ is the observation probability and $p(s_t\mid \hat{s}_{t-1})$ is the state transition probability.

For the observation probability, we develop a shape matching measure inspired by chamfer matching[16], a popular fast shape matching technique. In our matching method, the last detected head ellipse T is the "feature template", and the current body silhouette contour I is the "feature image". The original chamfer matching approach does not consider shape incompleteness. However, in our problem, the head contour may be missing in the joints position with the neck or in the possible intersection position with the arm when it is lifted above the head. So, we introduce a robust

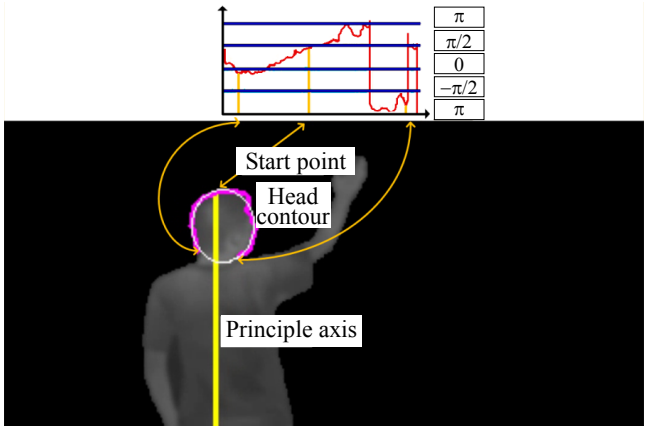


Fig. 6 A detected head showing the segmented principal axis of the human body and head contour with the gradient angle curve. The arrow lines indicate the start point, the left endpoint, the right endpoint of the head contour, and their corresponding positions in the gradient angle curve.

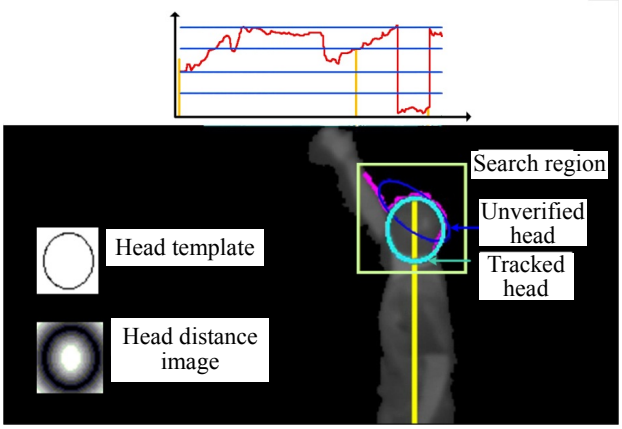


Fig. 7 Head tracking illustration in which the head contour is wrongly segmented due to distraction by the arm; hence, the fitted head is unverified. The last detected head is used as a template, and shape matching is employed to track the head. The head template and its distance transform image are also shown.

matching measure that detects matched points with a distance greater than a threshold as outliers and rejects them. We also apply the distance transform to the template image (see Fig. 7 for the head template and distance image) rather than the feature image for convenience. The proposed confidence measure is

$$P(o_t \mid s_t) = \omega \sum_{\substack{i \in I(s_t), \\ d_I(i) \leqslant d_{\text{th}}}} \frac{d_{\text{th}} - d_T(i)}{N_T d_{\text{th}}} + (1-\omega)\frac{N_T}{P_T} \quad (5)$$

$$d_{\text{th}} = \beta(a+b) \quad (6)$$

where the first term of Eq. (5) measures the matching confidence, with $d_T(i)$ denoting the chamfer distance between contour point i in the hypothesis image patch $I(s_t)$ and the closest edge point in T, d_{th} denoting the distance threshold, and N_T denoting the number of inlying matches. The second term measures the matching completeness, with P_T denoting the perimeter of the head ellipse in the template and ω denoting the weighting factor between the two terms. d_{th} is set in proportion to the head template's semi-axes a and b, and the coefficient β is set to be 0.1.

For the motion model, since the head is near the principle axis x_{p} horizontally and obeys the motion continuity vertically, the state transition probability is defined as follows:

$$P(s_t \mid \hat{s}_{t-1}) = \mathcal{N}(x_t - x_{\text{p}} \mid 0, \sigma_x^2)\mathcal{N}(y_t - \hat{y}_{t-1} \mid 0, \sigma_y^2) \quad (7)$$

where $\mathcal{N}$ is the normal distribution; σ_x^2 and σ_y^2 represent the covariance in the x and y directions, respectively.

The hypothesis space is generated in a neighborhood of $(x_{\text{p}}, \hat{y}_{t-1})$, and the hypothesis with the highest probability is obtained by exhaustive search. Figure 7 gives a head tracking example.

2.3 Hand localization

Compared to the head, localization of the torso and the arm region is relatively easy, and Fig. 8 illustrates the results for two different poses. The torso is modeled as a rectangle, whose height extends from the bottom of the head to the bottom of the body bounding box. The horizontal histogram of human body pixels below the head is calculated and binarized with a threshold. Then, the longest 1 segment is localized as the torso rectangle's width. Apart from the head and torso, only the arm remained, which is localized as the largest component with connected component analysis.

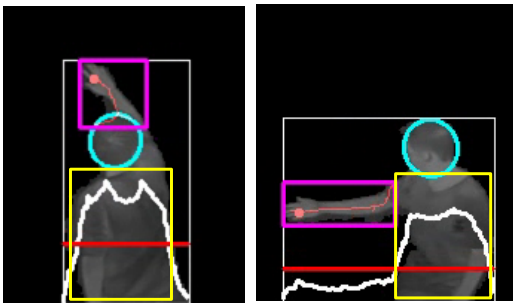


Fig. 8 The localized head, torso, arm, arm skeleton, and the hand point for two different poses. The horizontal histogram without the head and the threshold for locating the torso are also shown.

The identified arm region is skeletonized with the thinning algorithm in Ref. [17], and two endpoints of the skeleton are extracted. Based on the layout of the head, torso, and arm, the endpoint far away from the body is identified as the initial hand point. Since the skeleton is easily affected by local noises, we further extracted the initial hand point's neighborhood with the hand size estimated from the head size. The centroid of body pixels in the neighborhood is computed as the final hand point.

3 Gesture Recognition

The moving trajectory $\{\boldsymbol{p}_i = (x_i, y_i)\}_{1 \leqslant i \leqslant N}$ of the hand consists of the localized hand point in each frame. Gesture trajectory is segmented with static start and end positions by keeping the hand stationary for several frames[18]. Features are extracted from the gesture trajectory for recognition. Center of gravity $\hat{\boldsymbol{p}} = (\hat{x}, \hat{y})$ of the segmented trajectory is calculated by

$$\hat{\boldsymbol{p}} = \frac{1}{N}\sum_{i=0}^{N} \boldsymbol{p}_i \quad (8)$$

The distance feature l_i is calculated by

$$L_i = \| \boldsymbol{p}_i - \hat{\boldsymbol{p}} \|_2 = \sqrt{(x_i - \hat{x})^2 + (y_i - \hat{y})^2} \quad (9)$$

$$l_i = \frac{L_i}{L_{\max}}, \quad L_{\max} = \max_{1 \leqslant i \leqslant N} L_i.$$

The orientation feature θ_i is calculated by

$$\theta_i = \tan^{-1}\frac{y_i - \hat{y}}{x_i - \hat{x}} \quad (10)$$

In the recognition, we used a rule-based approach rather than a learning-based approach (e.g., an Hidden Markov Model (HMM[19]) or a Finite-State Machine (FSM[20])). Rule-based approaches are simple, fast, and

require no offline training. The rules are defined as follows for the predefined three gestures:

Lining The distance features decrease and then increase linearly, and the orientation features have two opposite values.

Circling The distance features are constant, and the orientation features have a 2π value range and increase or decrease monotonically.

Pointing The distance features are near to zero.

4 Gesture-Based Interaction

After a gesture is recognized, the interaction region needs to be mapped from the camera to the original contents. Camera projector calibration is employed to accomplish the mapping.

4.1 Camera projector calibration

The planar correspondence of pixels in the projection screen between the camera and the original projected contents is determined by a 3×3 homography matrix $\boldsymbol{H}$ as follows:

$$[sx_{\mathrm{p}} \quad sy_{\mathrm{p}} \quad s]^{\mathrm{T}} = \boldsymbol{H}[x_{\mathrm{c}} \quad y_{\mathrm{c}} \quad 1]^{\mathrm{T}} \tag{11}$$

where $(x_{\mathrm{c}}, y_{\mathrm{c}})$ is the projection screen's point observed in the image, and $(x_{\mathrm{p}}, y_{\mathrm{p}})$ is the corresponding point in the original contents. At least four such correspondence pairs are required in order to solve $\boldsymbol{H}$. Directly finding such pairs with the thermal camera is impossible because of the lack of texture of the imaging. We therefore develop a calibration method in two steps aided by the Web camera: $\boldsymbol{H} = \boldsymbol{H}_{2\mathrm{cam}} \times \boldsymbol{H}_{2\mathrm{cont}}$, where $\boldsymbol{H}_{2\mathrm{cam}}$ is the homography matrix from the thermal camera to the Web camera, and $\boldsymbol{H}_{2\mathrm{cont}}$ is the homography matrix from the Web camera to the projected contents.

$\boldsymbol{H}_{2\mathrm{cam}}$ is calculated in the first step. The four vertices of the projection screen in both cameras are detected by finding lines using the Hough transform[21] and obtaining the quadrilateral. The vertices in both cameras are used to calculate $\boldsymbol{H}_{2\mathrm{cam}}$. The results are illustrated in Fig. 9.

To get the homography matrix from the Web camera to the projected contents, we employed SIFT keypoints[12] to detect candidate correspondence points. SIFT points are scale-space extrema detected in a set of difference-of-Gaussian images in the scale space. Each keypoint is assigned a location, scale, orientation, and a descriptor; it is thereby invariant to scale and rotation and robust to illumination change and viewpoint change. Given the detected SIFT points in two images, the nearest neighbor matching algorithm from Ref. [12] is used. After this matching step, the corresponding point pairs are only putative. We use RANdom SAmple Consensus (RANSAC)[17] to impose the homography constraints on the previous matched point pairs and to find the correct correspondences, which are finally used to calculate $\boldsymbol{H}_{2\mathrm{cont}}$. The whole process is illustrated in Fig. 10.

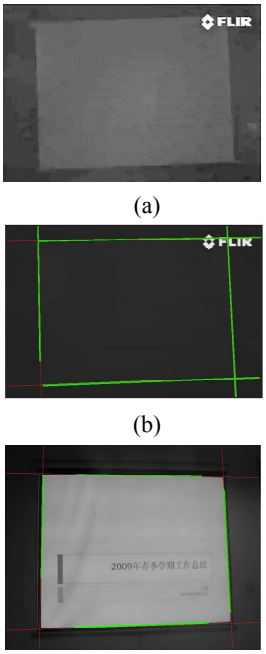
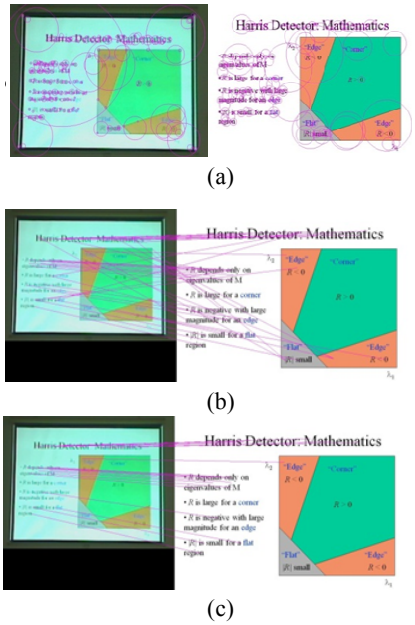


Fig. 9 Vertices detection with the Hough transform: (a) contrast-adjusted thermal image for viewing the screen better; (b) corner detection in thermal camera images; and (c) corner detection in Web camera images

Finally, we obtained the final homography by $\boldsymbol{H}=\boldsymbol{H}_{2\mathrm{cam}}\times\boldsymbol{H}_{2\mathrm{cont}}$. Though the image quality of the Web camera was not very good and no special pattern was projected for calibration, the experiment demonstrated the effectiveness of our method: it had a maximum pixel error of less than 8, which was sufficiently precise for the interaction.

4.2 Interaction

Upon identifying the interaction region of the recognized gesture with the calibration, the semantic interaction is carried out. In a PPT interactive presentation scenario, lining upward or downward controls the



(a) (b) (c)

Fig. 10 Illustration of finding the homography from the Web camera to the projected contents: (a) detected SIFT points in the visible camera and the original contents; (b) the putative matched point pairs; and (c) the final matched pairs by imposing homography constraints with RANSAC

forward or backward slide transitions, respectively; lining leftward or rightward horizontally makes annotations on the slide; circling controls the zoom in/out operations of the interaction regions in the slide; and pointing controls the activation of buttons in the slide. Once the gesture is detected and recognized, the corresponding actions can be performed in the PPT (such as annotation) and projected to the screen. Figures 11a and 11b illustrate a circling gesture and a lining gesture, respectively, in the presentation.

5 Performance Evaluation

We quantitatively evaluate the proposed system in terms of head and hand localization accuracy, gesture recognition rate, and computation time. The system runs on a Windows PC with an Intel Dual Core 3 GHz CPU and could be run at 20 frames per second in a single CPU core without any specific programming optimization. We conduct the experiments in online real-time scenarios with complex slide contents (e.g., colored figures and backgrounds).

(a) (b)

Fig. 11 (a) Circling gesture for zooming into the graph in the presentation; and (b) Lining gesture for annotating a PPT slide. The red annotation line is drawn on the original PPT slide and projected to the screen after the lining gesture is recognized.

The head and hand localization accuracy is key for the system performance and is defined as

$$\text{Localization accuracy} = \frac{\text{Number of correctly localized head (or hand)}}{\text{Number of total head (or hand)}}.$$

For the head, correct localization means that the overlap area of localized head ellipse D_l and ground truth D_{gt} must exceed 50%:

$$\text{OverlapRatio} = \frac{\text{area}(D_l \cap D_{gt})}{\text{area}(D_l \cup D_{gt})} > 0.5.$$

For the hand point, correct localization means that the point is localized in the actual hand region. The evaluation results are shown in Table 1, and Fig. 12 shows certain hand point localization results.

Table 2 shows the recognition rates of the three types of gestures. These results show that our system has achieved high recognition performance. The main error in recognition is caused by wrong gesture segmentation.

Table 1 Localization accuracy of the head and hand point

	Number of total	Number of correctly localized	Localization accuracy (%)
Head	6895	6839	99.2
Hand point	5092	5044	99.1

Fig. 12 Hand point localization results

Table 2 Gesture recognition results

Gesture	Number of gestures	Number recognized	Recognition rate (%)	Overall rate (%)
Lining	84	81	96.4	
Circling	52	50	96.2	96.7
Pointing	43	42	97.7	

6 Conclusions

In this study, we design a hand gesture based presentation system by integrating a thermal camera with a Web camera; this system provides a natural and convenient interaction interface for presentations. Quantitative experimental results show that the system performs efficiently. Currently, our gesture interaction is limited to one hand, and we plan to extend the work to handle two hands to enrich the system's interaction capability. The Web camera is only used for calibration, but additional information could be extracted for interaction, such as the shape of the hand (e.g., palm or fist). Also, other HCI applications such as gaming will be explored in future studies.

Acknowledgments

The authors would like to thank Nokia China Research Center for their support of this research.

References

[1] Erol A, Bebis G, Nicolescu M, et al. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 2007, **108**(1-2): 52-73.

[2] Wang F, Ngo C W, Pong T C. Simulating a smartboard by real-time gesture detection in lecture videos. *IEEE Transactions on Multimedia*, 2008, **10**(5): 926-935.

[3] Francke H, Ruiz-Del-Solar J, Verschae R. Real-time hand gesture detection and recognition using boosted classifiers and active learning. In: Proceedings of the 2nd Pacific Rim Conference on Advances in Image and Video Technology. Santiago, Chile, 2007: 533-547.

[4] Fang Y K, Wang K Q, Cheng J, et al. A real-time hand gesture recognition method. In: Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, Vols 1-5. Beijing, China, 2007: 995-998.

[5] Hu Z, Wang G, Lin X, et al. Recovery of upper body poses in static images based on joints detection. *Pattern Recognition Letters*, 2009, **30**(5): 503-512.

[6] Iwasawa S, Ebihara K, Ohya J, et al. Real-time estimation of human body posture from monocular thermal images. In: Proceedings of the 1997 IEEE Computer Vision and Pattern Recognition. San Juan, Puerto Rico, 1997: 15-20.

[7] Juang C F, Chang C M, Wu J R, et al. Computer vision-based human body segmentation and posture estimation. *Systems*, *Man and Cybernetics*, *Part A*: *Systems and Humans*, *IEEE Transactions on*, 2009, **39**(1): 119-133.

[8] Zou W, Li Y, Yuan K, et al. Real-time elliptical head contour detection under arbitrary pose and wide distance range. *Journal of Visual Communication and Image Representation*, 2009, **20**(3): 217-228.

[9] Sukthankar R, Stockton R G, Mullin M D. Smarter presentations: Exploiting homography in camera-projector systems. In: Proceedings of the Eighth IEEE International Conference on Computer Vision. Kauai, HI, USA, 2001: 247-253.

[10] Okatani T, Deguchi K. Autocalibration of a projector-camera system. *Pattern Analysis and Machine Intelligence*, *IEEE Transactions on*, 2005, **27**(12): 1845-1855.

[11] Wang F, Ngo C W, Pong T C. Lecture video enhancement and editing by integrating posture, gesture, and text. *IEEE Transactions on Multimedia*, 2007, **9**(2): 397-409.

[12] Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, **60**(2): 91-110.

[13] Stauffer C, Grimson W E L. Adaptive background mixture models for real-time tracking. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Ft. collins, CO, USA, 1999: 246-252.

[14] Chang F, Chen C J, Lu C J. A linear-time component-labeling algorithm using contour tracing technique. *Computer Vision and Image Understanding*, 2004, **93**(2): 206-220.

[15] Gander W, Golub G H, Strebel R. Least-squares fitting of circles and ellipses. *BIT*, 1994, **34**(4): 558-578.

[16] Barrow H G, Tenenbaum J M, Bolles R C, et al. Parametric correspondence and chamfer matching: Two new techniques for image matching. In: Proceedings of the 5th International Joint Conference on Artificial Intelligence. San Francisco, CA, USA, 1977: 659-663.

[17] Lam L, Lee S W, Suen C Y. Thinning methodologies — A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1992, **14**(9): 869-885.

[18] Davis J, Shah M. Visual gesture recognition. *IEE Proceedings-Vision Image and Signal Processing*, 1994, **141**(2): 101-106.

[19] Lee H K, Kim J H. An HMM-based threshold model approach for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999, **21**(10): 961-973.

[20] Bobick A F, Wilson A D. A state-based approach to the representation and recognition of gesture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, **19**(12): 1325-1337.

[21] Duda R O, Hart P E. Use of the Hough transformation to detect lines and curves in pictures. *Commun. ACM*, 1972, **15**(1): 11-15.