Hand and Foot Gesture Interaction for Handheld Devices

Zhihan Lv[†,‡] Muhammad Sikandar Lal Khan[‡] Shafiq Ur Réhman[†,‡*]

† Shenzhen Institutes of Advanced Technology(SIAT), Chinese Academy of Science, 518055, China.

‡ Immersive Interaction Lab(i2Lab),
Institutionen för Tillämpad Fysik och Elektronik, Umeå University, 90187 Umeå, Sweden.

ABSTRACT

In this paper we present hand and foot based immersive multimodal interaction approach for handheld devices. A smart phone based immersive football game is designed as a proof of concept. Our proposed method combines input modalities (i.e. hand & foot) and provides a coordinated output to both modalities along with audio and video. In this work, human foot gesture is detected and tracked using template matching method and Tracking-Learning-Detection (TLD) framework. We evaluated our system's usability through a user study in which we asked participants to evaluate proposed interaction method. Our preliminary evaluation demonstrates the efficiency and ease of use of proposed multimodal interaction approach.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

immersive multimodal interaction; smart phone games; foot gesture; HCI; mobile; vibrotactile.

1. INTRODUCTION

Recently there have been significant progress in multimodal human computer interaction(HCI) research due to advances in digital vision techniques and advanced sensor technologies. Modern HCI interfaces need not to rely on signal interaction modality (e.g. speech, touch and video only) or explicit input command from a single user [5]. The presence of computing power ranging from intelligent walls to hand held interfaces demand intuitive ways of interaction and human-centered design approaches. For effective human-machine communications researchers are considering different combinations of modality and multisensory approaches such as hand gesture [7], tooth clicks [11], eye blink [3] and foot based interaction [10] etc. Handheld devices (such as smart phones) are equipped with advance signal processing techniques and improved hardware, making real-time fusion of data and multimodal interaction possible.

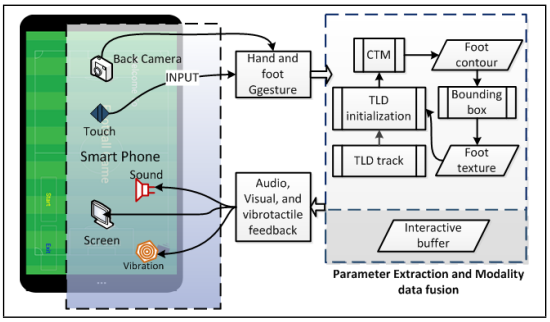

Figure 1: A block system diagram of hand and foot-gesture multimodal interaction for a smart phone.

We use feet along with hands in various everyday tasks (such as driving) and coordinated control mechanism of both modalities has been proven effectiveness in human computer interaction especially in virtual environments [12]. Considering handheld computing various foot gesture detection and tracking system have been proposed such as multitoe [1], kick [4], foot tapping [2] etc. Sangsuriyachot et al. [9] presented hand and foot gesture based interaction for tabletop environment. We believe that hand and foot gesture interaction can enhance immersive experience on handheld devices such as smart phone. To the best of our knowledge there has not been any study considering combination of hand and foot gesture for handheld devices.

In this study, we present a novel multimodal approach based on hand and foot interaction for smart phones (Fig. 1). The proof of concept is an immersive football game application on smart phone. For this game the user controls an augmented ball in virtual football field using hand and foot gestures on a smart phone screen. The foot gesture is detected and traced by smart phone camera sensor. The vibrotactile feedback is provided to user hands (through smart phone vibration) and to foot (using additive vibration sensor) along with visual and audio output. The primary contributions of this paper are

- An extended algorithm for foot detection and tracking for smart phone application without additive hardware.

*Corresponding author 'shafiq.urrehman@umu.se'

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502163.

Figure 2: Example frames of unexpected detected contour due to sudden motion. a) Correct foot-contour detection; and false foot-contour because of (b) the pattern on the shoe (when zoom in), (c) the bottom of a trouser when stretch out leg; (d)the knee when the foot is obscured

- A novel multimodal hand and foot gesture interaction platform for hand held devices.

2. IMMERSIVE FOOTBALL - A PROOF OF CONCEPT

To study the physical capabilities of user and effectiveness of proposed hand-foot based immersive multimodal interaction for smart phones, we develop a multimodal interaction game based on hand and foot interaction.

2.1 Enhanced Foot Gesture Tracking

In this work, we have extended algorithm presented in [8] for robust foot gesture tracking in real time using Tracking-Learning-Detection (TLD) framework [6].

One limitation of previous algorithm [8] is observed in the *'shoe target test'*; i.e. a visual inspection to check how good the algorithm has localized the users' shoe in each and every video frame. The false detection is noticed due to sudden appearance of foot like shape contours on nonfoot areas of smart phone's camera view such as bottom of a trouser, the part prominent pattern on the shoe, and/or even the knee shape (while sitting) as shown in Fig. 2. The motion of the most similar contour isn't equivalent to the motion of the foot exactly. Furthermore, the applications which need more accurate localization of the foot, such as kicking dynamic football, a minor foot-contour detection error can lead to undesired outcomes. Whereas in previously presented applications, such as foot piano, the motion of the most similar contour is classified as equivalent to the motion of the foot-contour. But in the foot game, if the other parts of body besides foot become interactive in the football game, the player will loose the control of the football. Secondly, the computation complexity of the contour based template detection approach is $O(M \times N \times L)$, where M and N are the width and height of the image, L is the length of the template curve. Even if it is enough to detect the foot-contour in the image in most case, it is not fast enough for tracking specified object in each and every frame. In practice, although the camera screen image is resized into about one quarter as (320 $\times$240), but still maximum performance can not be achieved for current multimodal application scenario using previous algorithm [8].

Our algorithm uses dynamic programming to localize the foot-contour and employ tracking, learning and detection (TLD) framework to strengthen the tracking process. TLD algorithm combines elements from tracking, learning and detection in the 2D image space to make it a long-term tracker [6]. The TLD tracker uses a tracking strategy of

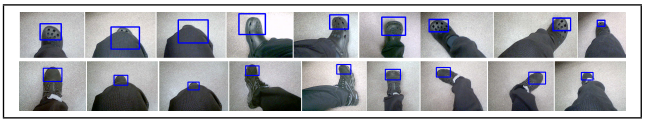

Figure 3: The tracking results for a variety of foot-gestures with either obvious (1st row) or weak features (2nd row).

the overlapping blocks, and tracks every block by Lucas-Kanade optical flow method. Hence it can only trace the user-specific foot texture image during tracking, but can't initialize by importing a generic foot model automatically. Even if the TLD always does the dynamic PN-learning process, the shoes with different appearance can't be detected by presetting only one shoe texture image initially and the shoes with totally different property such as different pattern or color can't be learned. Moreover, in TLD tracking algorithm, the off-line trained generic detector localizes object based on specific features and the online trained validator decides which object correspond to the object of interest. In case of the shoe, which may not have distinctive feature, the classification learning and training method such as cascade can't detect diverse shoes from a content-rich image. In view of this, the contour based template detection (CTD) approach is an indispensable step for identifying a variety of shoes in the beginning. CTD gives TLD the region of interest(ROI) for initialization. Our enhanced algorithm has following steps:

1 - Using CTD [8] detect and localize the user shoe, and create a bounding box BB ={ top-left(x^{min}, y^{min}), bottom-right(x^{max}, y^{max})}, where $x^{min}, y^{min}, x^{max}, y^{max}$ are the minimum and maximum values of the x and y in given frame;

2 - The bounding box is given to TLD algorithm as ROI; ROI={$(p^{11}, p^{12}, \cdots, p^{1M})$,$(p^{21}, p^{22}, \cdots, p^{2M})$, $\ldots$,$(p^{N1}, p^{N2}, \cdots, p^{NM})$}; where p represent the density of every point of the bounding box;

3 - TLD algorithm tracks the ROI from the camera view with the PN-learning process, and provides the information to the application layer;

The enhanced algorithm (see algo. 1) successfully solves all of the unexpected detection results of foot/shoe look-likes. Our algorithm execution time is faster and is accurate even when the foot/shoe is moving fast and in image blur cases. It's worth mentioning that this method can detect almost all obscured foot and almost all kinds of the foot-gesture in the football game, as long as the correct foot texture is detected in the first frame. As in Fig. 3, 1st row is the result of the shoe with obvious feature and 2nd row is the result of the shoe with weak feature. To improve the efficiency of the proposed algorithm, the size of first frame is scaled down to 25% and then CTD is applied. From the second frame the image is scaled to 12.5% and then TLD tracking is employed. The 'shoe target test' results indicate increase in both efficiency and accuracy success rate of the tracking. We have performed extensive test to evaluate the robustness of the foot detection and tracking algorithm on our own collection of foot-gesture video data-set. Algorithm

results are very much improved as compared to previous studies. The proposed algorithm successfully located and tracked the foot-gesture in almost all videos even when a sudden appearance of other objects covers most of foot. The overall foot-gesture recognition rate of 99% for various foot gestures and the real-time processed time consuming is only 18% of the previous studies.

Algorithm 1: The TLD algorithm with CTD localization

Input: ROI_0 , which is the ROI of the first frame, $L_0 = \{R_0\}$, where R_0 is the content surrounded by ROI_0 in the first frame; L_0 is the online model in time $t = 0$.

Output: current localization ROI_t

initial localization $\leftarrow ROI_0$;
current localization $\leftarrow 0$;
while *active input feed from camera in time* t **do**
  Tracking
    -Forward-Backward tracking by running Lucas-Kanade
    -Compute median flow
  Detection R in online model L_{t1}
    -Variance Filter
    -Ensemble Classifier
    -Nearest Neighbor Classifier
  Learning
    -$P - N$ Experts
      -$Lt \leftarrow L_{t1} \cup$ *Positive samples from growing.*
      -$Lt \leftarrow L_{t1} \setminus$ *Negative samples from pruning.*
  Integrator
    -Validation
    -$R_t \leftarrow$ *the most confident evaluated result*
  $ROI_t \leftarrow$ *bounding box of* R_t

2.2 Multimodal Football Game

We develop a real-time multimodal football game smart phone application. It renders the game graphics and game status information using smart phone's screen, audio and vibration. Augmented reality image rendering technology is employed. The players interact with the game using both foot gestures and finger touch on screen, which triggers the interaction event and generate activity sequences for interactive buffers. The trajectory of the football is changed by interactive buffers. In our system, the players use foot to kick the ball on the smart phone screen, meanwhile, they guard the goal post area using hand-finger. The game rules are as following:

1. The player interacting-foot must be placed in mobile camera view for 5 sec. for detection and localization.

2. The football always bounces back and forth on the whole screen area till end of the game and the players try to keep the football (using touch and kicks) under his/her control while preventing it going to goal-post with hands-gesture(i.e. touch on the smart phone screen).

3. The player get 1 score if she/he touches (or kicks) the football once, and the game is over if the football enters into the highlighted goalpost.

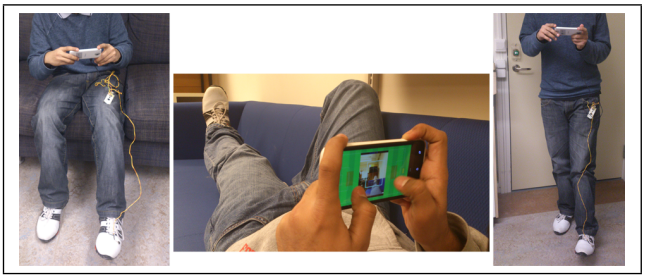

Figure 4: The players can play immersive football game application in a variety of postures; i.e., sitting, lying, and/or standing.

Questionnaire Items	Mean (M)	Standard Deviation (SD)
Playing the game was Interesting	2.14	0.69
Playing the game was Fun	3.05	0.89
Playing the game was Esay	3.15	1.02
Playing the game was Tiresome	3.31	1.15
Playing the game was Difficult	3.57	0.78

Table 1: Perceived suitability of immersive multimodal football using descriptive statistics before the game (fully agreed=5).

We also present the players with two types of vibrotactile rendering. Firstly, the football motion and collision with the player's finger is presented to user's hands through smart phone's built in vibration motors. Secondly, an additive vibrotactile rendering device is attached to the player's interacting foot; i.e., an extended vibration feedback to the foot kicking the ball. For this purpose an IOIO device is connected to the smart phone by USB cable; it is used for controlling a vibration motor attached to the interacting foot. The vibration feedback on foot renders a similar tactile sensation to the player, as if the player's foot is really kicking a football. The system not only merges the hand and foot interaction with the scene but also broadcasts information to the sight, auditory and vibrotactile directly. These modalities are integrated into a small smart phone, so the players can play immersive football game in variety gestures, for example standing, sitting, lying as shown in Figure 4.

3. PRELIMINARY USER STUDY AND RESULTS

There were 12 subjects aging from 25 to 35 in total involved in the preliminary user study. All the participants had smart phones and had prior experience with mobile sensors such as camera, vibration etc. The motivation of our user test had been clearly explained to all the participants. First, we introduced the purpose of our experiments to the participants and explained how to use the GUI and application scenarios. The game rules were also explained to the users. Each participant held the smart phone in both hands such that the camera's eye could view the user's lower leg. The users were asked to perform as they like i.e, sitting, standing and or lying. 90% of the participants preferred playing game while siting. In order to evaluate the popularizing and players mood values, we used PANAS ques-

Questionnaire Items	Mean (M)	Standard Deviation (SD)
Playing the game was Interesting	4.42	0.79
Playing the game was Fun	4.52	0.89
Playing the game was Esay	3.85	0.69
Playing the game was Tiresome	2.88	0.59
Playing the game was Difficult	2.77	0.95

Table 2: Suitability of immersive multimodal football using descriptive statistics after the game (fully agreed=5).

tionnaire [13] before and after the game. For PANAS questionnaire analysis we used paired sample t-test. A useful measure of user satisfaction can be made if the system evaluation measure is based on observations of user attitudes towards the system. Thus, it is possible to measure user attitudes using a questionnaire, e.g., "Is this application interesting? 1 very boring (negative) to 5 very interesting (positive)". The satisfaction levels of our experiments were measured by giving participants questionnaires.

There was significant increase in overall positive affect after playing multimodal football game app. on smart phone. After the game M=3.79 , SD= 0.85 as compared to before the game M= 2.56, SD= 0.69 with $t_{11} = -3.99$ and $p = 0.001$. However very slight change was observed in negative affect before and after the game; i.e., after M= 1.35, SD= 0.39 and before M= 1.57, SD= 0.49. The questionnaire showed that the designed multimodal football game was perceived easy and fun as shown in Table 1 and 2. It was also noted from the participants remarks that the game had positive affect on their alertness. It was also mentioned that the vibrotactile feedback on foot improved their experience and increased excitement. The technical implementation of the game also showed successful performance. The participants were able to perform all desired foot gestures.

In a real world football game, the players kick the football in various direction using different foot postures. The approach proposed in this paper can only get the 2D location of the foot on the ground plane. Although the approach can track foot-gesture with different 2D-postures, it can't get all of the exact 3D angle parameters. These issues can be solved using additive sensor devices and/or software algorithm modification. We do not have any measure to quantify the performance of the user; so cognitive workload questions can not be answered in this work.

4. CONCLUSIONS AND FUTURE WORK

In this study, we propose an immersive multimodal interaction game on smart phone, with focus on the detection and tracking of the foot-gesture and hands in real time. Based on the optimized interaction approach, we combine both foot and hand gestures as input and render game feedback to the users using vibrotactile patterns (both on hands and foot), sound and vision. Through the analysis of the user study results, it is found that the combination of multiple input and output is an intuitive method for novice users and combination of modalities can provide an enhanced immersive experience. It allows players to mobilize more body part for multitasking interaction. In future studies, we will consider hardware and software modification to extract the foot-gesture orientation (i.e., yaw, pith, roll) parameter. We believe that it's entirely possible that the multi-player competition for the football by the real foot interaction can be rendered on the smart phone in the near future. To evaluate the multi-dimensional tracking approach, we will consider a real-time 3D football-kicking application as the interactive framework.

Acknowledgments.

The authors are thankful to the Chinese Academy of Sciences Fellowship for Young Foreign Scientists (2012Y1GA0002) and National Natural Science Fund of China (61070147).

5. REFERENCES

[1] T. Augsten, K. Kaefer, R. Meusel, C. Fetzer, D. Kanitz, T. Stoff, T. Becker, C. Holz, and P. Baudisch. Multitoe: high-precision interaction with back-projected floors based on high-resolution multi-touch input. In *23rd Annual ACM Sym. UIST*, 2010.

[2] A. Crossan, S. Brewster, and A. Ng. Foot Tapping for Mobile Interaction. In *24th BCS Conf. HCI, UK*, 2010.

[3] K. Grauman, M. Betke, J. Gips, and G. R. Bradski. Communication Via Eye Blinks - Detection and Duration Analysis in Real Time. In *IEEE Conf. Comp. V. and P. Recognition*, 2001.

[4] T. Han, J. Alexander, A. Karnik, P. Irani, and S. Subramanian. Kick: Investigating the Use of Kick Gestures for Mobile Interactions. In *13th Int. Conf. HCI with Mobile Devices and Services*, 2011.

[5] A. Jaimes and N. Sebe. Multimodal Human computer Interaction: A survey. *Comp. V. and Im. Understanding*, 108:116–134, 2007.

[6] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking Learning Detection. *IEEE Trans. on Pattern A. and M. Int.(TPAMI)*, 6(1), 2010.

[7] P. Mistry and P. Maes. SixthSense: a wearable gestural interface. In *ACM SIGGRAPH ASIA 2009 Sketches*, 2009.

[8] S. Réhman, A. Khan, and H. Li. Interactive Feet for Mobile Immersive Interaction. In *ACM Int. Workshop MobiVis Workshop at MobileHCI*, 2012.

[9] N. Sangsuriyachot and M. Sugimoto. Novel Interaction Techniques based on a Combination of Hand and Foot Gestures in Tabletop Environments. In *10th Asia Pacific Conf. HCI*, 2012.

[10] J.-H. Seo, J.-Y. Yang, and D.-S. Kwon. Laser scanner based Foot Motion Detection for Intuitive Robot User Interface System. In *IEEE RO-MAN*, 2012.

[11] T. Simpson, M. Gauthier, and A. Prochazka. Evaluation of Tooth-Click triggering and Speech Recognition in Assistive Technology for Computer Access. *Neurorehabil Neural Repair*, 24:188–194, 2010.

[12] D. Valkov, F. Steinicke, G. Bruder, and K. Hinrichs. Traveling in 3D Virtual Environments with Foot Gestures and a Multi-Touch enabled WIM. In *Int. Conf. Virtual Reality*, 2010.

[13] D. Watson and L. A. Clark. Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS scales. *J. Personality and Social Psychology*, 54(6):1063–1070, 1988.