Interactive Feet for Mobile Immersive Interaction

Shafiq ur Réhman
Immersive Interaction Lab.(i2Lab),
Department of Applied Physics and Electronics,
Umeå University Sweden, 901 87 Sweden.
shafiq.urrehman@tfe.umu.se

Abdullah Khan
Umeå University Sweden, 901 87 Sweden.
mcs10akn@cs.umu.se

Habio Li ∗, ‡
∗ College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 21000, China.
‡ School of Computer Science and Communication, Royal Institute of Technology (KTH), 100 44 Sweden.
haiboli@kth.se

MobiVis Workshop at MobileHCI'12, September 21–24, 2012, San Francisco, CA, USA.

Abstract
In this paper we propose a novel algorithm for foot-gesture tracking method in mobile phones. To evaluate our proposed algorithm we develop two application scenarios for mobile immersive interaction experience based on audio, vibrotactile and foot interactions. For current studies we have located and tracked foot-gesture using template matching algorithm. The strength of proposed algorithm is demonstrated based on a successful completion of the given tasks. In the first application scenario the user is presented with an *immersive fun dialing*, i.e. dialing desired phone numbers using foot-gestures, while in the second application scenario, the user is provided with an *immersive music game* for unlocking keypad using foot-gesture on a smart phone. Our algorithm not only successfully locates and tracks foot-gesture but also can detect and track shoe of any size. These studies show the effectiveness of foot-gesture on mobile phones in real life situations.

Keywords
immersive interaction; vibrotactile rendering; foot-gesture; mobile interface; mobile HCI.

Introduction
In today's world the usage of cellular phones has become a part of daily life, what is especially visible from the number of were 6 billion mobile subscriptions reaching 6 billion at the end of 2011 (i.e. 87 percent of the world population) [1]. Nowadays smart phones have greater processing

power than ever before, what is revolutionizing the use of cellular technology, making them more accessible for human mobile interaction [9]. Interactive user interfaces and high portability have increased the usage of mobile devices and Tablet PCs more than ever. User input on a mobile device is usually done by a touch screen or through buttons while output relies mostly on the visual feedback [10, 15]. Current smart phones are equipped with integrated cameras which can be used as the sensory input devices because mobiles have greater connectivity and can be taken almost anywhere, hence supports the concept of ubiquitous/pervasive and immersive computing. Besides this, interaction through camera input by gesture detection provides third dimension for interaction while input through human gesture detection on smart phones make keypad and touch screens redundant. Experiments reveal 3-D space increases by 5-10 percent of interaction resolution of 2-D space of touch screens [15].

The introduction of camera and other sensors in smart phones paved new ways for mobile-human immersive interaction such as usage of audio, visual, gesture based and/or vibrotactile interaction [7], [3]. At some instances semi-hand free interaction with smart phones become inevitable, especially when human hands are dirty or engaged in other activities. One interesting approach for the semi hand and/or immersive mobile interaction can be using foot gestures as an input method. Interaction based on human gestures is more natural for human beings and researchers have explored intuitive interaction-method in mobile computing. The foot gesture localization has been popular in Virtual and/or Augmented Reality application, e.g.; Cyberboot [2], and WIM (World-in-Miniature) [12]. Recently the foot gesture is considered for controlling smart phones movements [3], selecting menus [10] and playing the games [6]. In this paper, we demonstrate immersive mobile interaction prototype methods based on foot gestures. The concept has been suggested by [6] and recently explored by [5], but both work have used an extra mounted cameras or external Xbox Kinect camera for foot-gesture detection and tracking.

Our work proposes an optimized template matching algorithm for foot-gesture detection and tracking for smart phones which needs no add up hardware. For mobile immersive interaction, the users hold the mobile device in hands with clear screen-view while complement/additive information is provided using audio, vibrotactile and foot-gesture interactions. In order to show the feasibility and find out technical issues for proposed audio, vibrotactile and foot-gesture based mobile immersive interaction, two application scenarios have been designed; namely '*immersive fun dialing*' and '*immersive music game*' for unlocking keypad. The application scenarios are developed on Android SDK initially but proposed algorithm can also be extended to other smart phone platforms. For these application scenarios, Augmented Reality (AR) rendering is employed to show interactive visualizations while providing an immersive and user-friendly interface.

Mobile Immersive Interaction

In proposed method, the object of interest (i.e. foot-gesture) is extracted and tracked from a smart phone camera video. We have estimated foot motion parameters (such as speed and direction) which are used to detect collision with the augmented objects on smart phone's screen. To provide a user an interactive and immersive interaction, the extracted parameters information are rendered using audio, video and haptic rendering (as shown in Fig. 1); i.e., immersive interaction through mobile screen, speakers and vibrotactile motor fitted in a mobile phone.

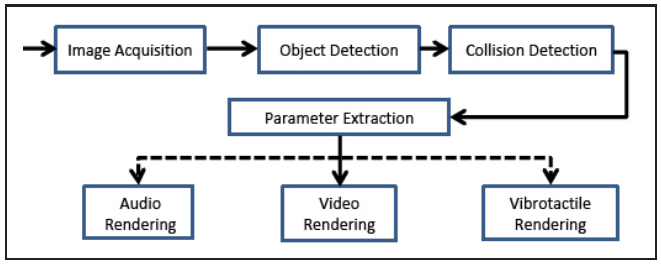

Figure 1: A block system diagram of foot-gesture detection and rendering for Mobile Immersive Interaction.

Foot Gesture Detection

One problem in applying current computer vision and image processing techniques on smart phones is that most of the algorithms are designed and developed on an assumption that the camera-position is stationary with a clear view of object to be tracked. In case of smart phones the camera is hardly ever stationary and it can be at any distance from the object of interest. Moreover, problem becomes more complex when the object of interest is also in motion. Another issue in object segmentation and tracking is relatively less computing resources available to smart phones as compared to PCs. This lowers the frame rate rendering which ultimately leads to a bad human-smart phone interaction. Researchers have modified some object and pattern recognition algorithms to perform human-smart phone interaction in real time which can differentiate between objects but do not tell what this object actually is [11, 13]. Here we propose an optimal template matching based object segmentation and tracking algorithm which considers the possibility of variable size of objects and its motions. In our case the template is a boundary curve around the object of interest, i.e. a grey scale boundary curve that can have higher or lower number of points (pixels) as compared to edge image which include only the boundary points of the object [8]. First, a human foot template is modeled off-line and then it is compared with image frames acquired from smart phone's camera to find an optimal match. Formally, the foot-template has size k with pixels $p1, p2, p3...pk$ and it is divided into groups called segments S $= \{$s1, s2, s3 $\cdots$ sN$\}$, where N is the total number of segments in a template. The length of each segment is fixed up to 3 boundary pixels which gives 33% deformation between two consecutive segments, hence allowing the whole template to deform. To search each segment in a given image we regarded this task as a problem of optimality, which is solved by dynamic programming. Considering deformation $D = \{(x1, y1), (x2, y2)...(xN, yN)\}$, where x and y represent the coordinates of the segments and converting a given image of size m x n into 1-D column matrix Cx; where $x = \{1, 2, ..., N\}$ is the spatial representation of image-pixels. The pixels in each column is given as $IP = \{$ip1, ip2 ... ipmn$\}$. The pixels of template in xth segment can be given by $pcx...p(cx + 1) - 1$ and Cx by equation;

$$Cx = 1; \qquad for \quad x = 1;$$

and

$$Cx = min(j = cx - 1, ..., k : |pj - p(cx - 1)| > 3); \\ for \quad 2 <= x <= N$$

where $|pj - p(cx - 1)|$ is the Euclidean distance between two boundary pixels. First template segment is considered and matched with pixels in C1 and the number of matches is calculated. Similarly second segment is matched with C2 and so on. The number of matches at each pixel is regarded as weight W(k,x) where $k = \{1, 2, 3, ..., mn\}$. Also the pixel in column Cx can connect with pixels in C(x+1) only if it satisfies the following equation $|ip(k, x) - ip(k, x + 1)| = 1/3|(pcx - p(cx + 1))$. This results in a weighted graph

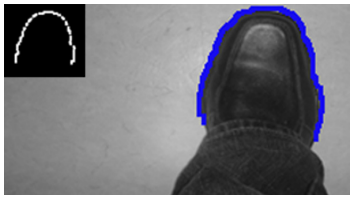

Figure 2: Foot template and a "detected-foot" is in a given frame.

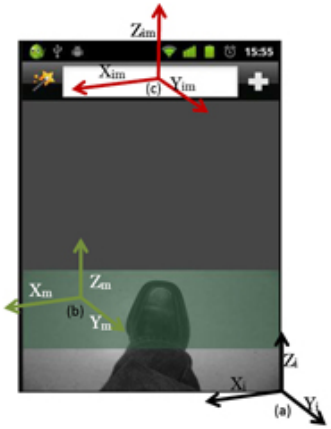

Figure 3: Co-ordinate systems description used by our foot-gesture detection algorithm for collision detection in Mobile Immersive Interaction.

commonly known as *'trellis'*. If a path from pixel Wp(k,1) in C1 to pixel Wq(k,xq) in Cx for $1 < xq < N$ is an optimal path then dividing the path at any point Wr(k,xr) $1 < xr < xq$ in the optimal path; the path (Wp - Wr) and (Wr - Wq) should also be the optimal paths. The final weight for each pixel is termed as accumulated weight AW(k,x). It can be given with the equation

$$AWq(k,x) = Wq(k,x) + max(Wr(k,x-1))$$

In trellis, each cell in column contains an accumulated weight as well as the index of the matched pixel of previous segment thus facilitating back tracking. Once the algorithm reaches the right most column CN; the cell with maximum weight is found. From this cell our algorithm starts back-tracking. The optimal path obtained by back-tracking is the true deformation of template as shown in Fig 2. The position of human-foot from the coordinate system of image frame is extracted; i.e., (X_i, Y_i, Z_i); where Z_i gives the directional information. It is translated to the smart phone's rendering-resolution coordinate points (i.e. (X_m, Y_m, Z_m)) to detect the collision. After collision detection the augmented object is moved to a known location; i.e., (X_{im}, Y_{im}, Z_{im}) as shown (red) in Fig. 3.

To verify the accuracy of our algorithm we have tested it on our dataset; which contained 120 mobile video sequences (more than 800 frames each) of single foot-gesture for various sizes. Our dataset contained video-sequences from clustered office as well as outdoor lighting conditions. The proposed algorithm successfully located and tracked the foot-gesture in almost all videos. The error in foot-gesture tip tracking is only noticed during a sudden appearance of other objects (such as occlusion due to clothes) on top of object of interest (i.e. foot). The overall foot-gesture recognition rate of 98% is noted for private video sequences showing various shoe-shapes with sudden state change (see Fig.4).

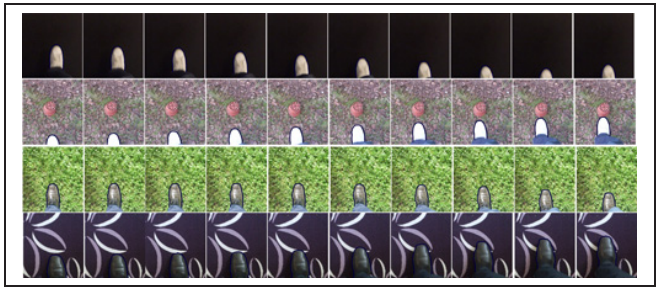

Figure 4: Results for template based foot-gesture tracking algorithm applied on our dataset.

Interactive Feet and Mobile Immersive Interaction

We proposed three layer architecture for a mobile immersive interaction application scenario based on foot-gesture. First layer is the GUI layer which is used for augmented object rendering. The second layer performs communication between upper and lower layers and renders collected information using audio, video and/or vibrotactile rendering schemes. The functionality of the third layer is common for both application scenarios. It is vision module layer that deals with detecting the foot-gesture in the video frames (i.e. vision module layer). However it is worth mentioning that captured frame is resized to 320 x 240 keeping in mind the limited computing resources available in smart phones.

For human foot detection and tracking, we considered only a small window of a given frame; i.e., region of interest (ROI). Once the foot is detected ROI changes its location based on the movement and current position of the human foot. Following two application scenarios are developed to

investigate how efficient the proposed method is for immersive interaction in smart phones.

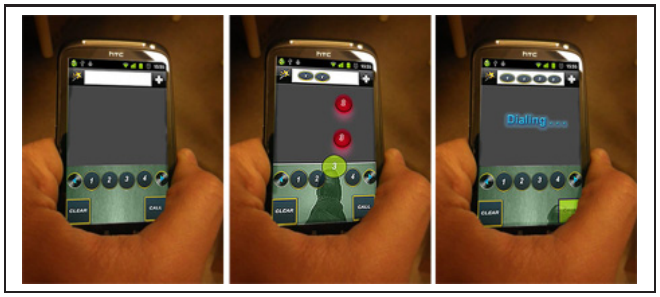

Figure 5: Foot-gesture based *'Immersive Fun Dialing'* application scenario. The user kicks the augmented digits and then presses the dial-key area using foot-gesture.

Scenario I: Immersive Fun Dialing
We have developed a keypad dialing application: i.e., 'Immersive Fun Dialing' on Android platform. Fun dialing is a foot-gesture based smart phone key-pad. The user presses the augmented digits on mobile-screen using foot-gesture from the mobile phone's camera-view. Augmented keypad is designed with a set of 4 digit-keys rendered at one instant on mobile screen. To access the required augmented-digits, when the user wants to select any digit on mobile phones display one can slide his/her foot in front of camera-view. The digit-key can be selected by pressing/hitting the specific key using foot-gesture, until it starts highlighting/moving (see Fig. 5, Fig. 6 for explanation).

We have used three-layer architecture for simplicity. The first layer, which renders the augmented digit-keypad, is developed using Android SDK. The 2nd-layer is responsible for the communication (between the 1st and 3rd-layer) and performs actions based on extracted information from both layers and rendering content using vibration and/or audio. The frames are acquired by the built-in camera in YUV color space. The captured frames are processed and converted into an image-size which is supported by the smart phone's display-hardware. After this the processed frame is passed to the 3rd-layer for detecting foot-gesture and parameter estimation. Simultaneously, the processed image with marked foot-position is sent back for visual rendering, so the user can see the movement of his/her foot. The complete GUI renders the camera-view with augmented graphical shapes and renders audio and vibrotactile information for more intuitive immersive interaction experience as shown in Fig. 5. The connection layer (i.e. 2nd-layer) uses android telephony to perform the make-call/dialing function. It is worth mentioning that the coordinates of camera frame differs according to the smart phone's hardware. The coordinates of the foot-gesture in each frame are taken and translated into display's coordinates. When coordinates of any key in the GUI display coordinate system become equal to the translated coordinates of human foot, a collision is detected.

Scenario II: Immersive Music Game
In order to check the feasibility for immersive mobile interaction using foot gesture, we have developed second application scenario; namely 'immersive music game' for unlocking. This application gives an option to unlock certain application and/or keypad on the user's android based smart phone using foot-gesture, what acts as a user defined music-pattern password. Once the music-pattern is selected by the user, the application runs in a background to check if application locked by the user is touched. If the locked application is touched the 'foot-gesture based immersive music application' comes to foreground and asks for a special music-pattern. The application uses SQLite databases for storing the user defined music-pattern. When the volume up button is pressed the application starts to

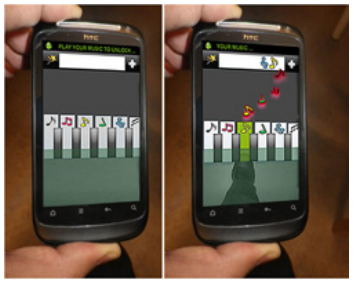

Figure 6: Music game for unlocking smart phone's keypad using a foot-gesture.

look for the movement of the foot through android camera as shown in Fig. 6. When a collision is detected with any of the augmented piano-key, the code related to that key is generated by the program. If the code generated is equal to the code stored in the database, the application/touch-pad gets unlocked. This application works on the above mentioned three layer structure. The first layer displays augmented graphics where each piano-button has an ascribed specific code. The second layer is responsible for communication between upper and lower layer as well as handles database for pattern matching.

User Study: Participants and Procedure

Participants were recruited from the campus of Umeå University, Sweden, including both students and staff members. There were 15 people aging from 20 to 50 in total involved in the experiments. The age of 80% of them is from 26 to 35 years. All the participants had smart phones and had prior experience with mobile sensors such as camera, vibration etc. For each application scenario all the users were provided with one training session up to 1-5 minutes. 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. Each participant held the smart phone such that the camera's eye could view the user's lower leg. To obtain sufficient experimental data, the users were asked to perform two sessions both sitting and standing. They were also asked to use each application 10 times in one session and to point out their "experience" and "unsuccessful-attempts". In the end, each participant was given a questionnaire. Each question was elaborated and discussed with the participants if required.

Results and Discussion

The immersive mobile interactions based on foot-gesture were considered intuitive and overall feedback was positive. Both application scenarios provided real-time interactions; i.e. processing/rendering time was 20-25 fps. For 'Immersive Fun Dialing' users found vibrotactile feedback along with visual feedback very helpful. Moreover, the user's interaction (audio and visual information augmentation) was increased by additive vibrotactile pattern as confiremed by all the users. The usability and the user experience is vital for any mobile phone application, but usability evaluation in above mentioned application scenarios is a challenging issue due to the uniqueness of the features and its novelty [14]. For this work, usability is defined according to how well the applications can be used to achieve the goals with effectiveness, efficiency and satisfaction [4]; i.e., in our case goals were smart phone keypad dialing task and smart phone keypad unlocking task. To evaluate our foot-gesture based immersive mobile interaction effectiveness we have used the application dependent implicit criteria; i.e. criteria based on the argument that it is difficult in practice to obtain a 'true' measure, thus the results are evaluated as an output of a complete system. The reported accuracy is compared against unsuccessful attempts. The percentage accuracy is shown in table 1.

Application Scenario	Successful (%)	UnSuccessful(%)
Immersive Fun Dialing	90	10
Immersive Music Game	95	5

Table 1: Effectiveness of foot gesture based mobile immersive interaction.

The results of these experiments showed that the performance of almost all the subjects is increased with the pas-

sage of time, i.e. less unsuccessful attempts, indicating that user's perception of the system functionality (i.e. user-training) has increased with training and so did its usability. Efficiency is measured as how much effort is required in order to accomplish the task [4]. In the experiments, we used the 'total completion time' taken for a given task by the user as an indicator of efficiency. The task completion time was measured by computing a delay between the time when a user 'started' an application by pressing side 'camera-button' and the time when the user pressed a 'dial/play-area' in GUI (see Fig. 5). The delay time contained three parts: *rendering time*, *cognitive time* and *move-and-press time*. Since the rendering time was approximately constant, we assumed that the delay time was a reliable indicator of 'cognitive time' plus 'move-foot-and-press'. The delay time with correct results in both sessions decreased, what can be seen in table 2. For both cases, the delay declined by 23.95% and 10.89%, with 14.17% drop of overall delay, implying that after training the users tended to take less time to perform the tasks. A careful examination of the delay results revealed that there was no stable pattern in delay time when a user made mistakes. Furthermore, user Studies have shown that the foot wearing shoe of any size and shape is detectable at certain distances as long as it is in the camera view of smart phone and the results are rendered in real time.

Mean (s)	session 1	session 2
overall delay	0.73705	0.59535
Immersive Fun Dialing delay	0.8568	0.6173
Immersive Music Game delay	0.6823	0.5734

Table 2: Efficiency measured for foot-gesture based mobile immersive interaction.

Concluding Remarks

In this paper, the real-time application platforms for designing a mobile immersive interaction experience based on foot gesture, audio and vibrotactile interaction for smart phone is presented. It proposes a foot-gesture tracking algorithm based on template matching for smart phone applications. The performance of foot-gesture tracker along with its robustness is measured indirectly, using application success rate based approach. The user tests show that after a little training subjects are able to enjoy both application scenarios, hence the current work can also be used in other mobile immersive application such as gaming and entertainment industry. We feel however, that further efforts are needed to achieve better performance in the vision module and therefore we plan to enhance the software and add some hardware to our application scenarios in the following studies. From the applications discussed above it can be concluded that mobile phones, along with communication platforms, can be used as creative and immersive interaction devices.

References

[1] International Telecommunication Union, Key Global Telecom Indicators for the World Telecommunication Service Sector, 2011, 2011.

[2] Choi, I., and Ricci, C. Foot-Mounted Gesture Detection and its Application in Virtual Environments. In *IEEE Int. Conf. on System, Man and Cyber-netics* (1997), 4248 – 4253.

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

[4] Faulkner, X. *Usability Engineering*. Palgrave Macmillan, 2002.

[5] Han, T., Alexander, J., Karnik, A., Irani, P., and Subramanian, S. Kick: Investigating the Use of Kick Ges-

tures for Mobile Interactions. In *13th Int. Conf. HCI with Mobile Devices and Services (MobileHCI 2011), Stockholm, Sweden* (2011).

[6] Paelke, V., Reimann, C., and Stichling, D. Foot-based Mobile Interaction with Games. In *ACM SIGCHI Int Conf Adv Comp Ent Tech (ACE'04), Singapore* (2004).

[7] Réhman, S., and Li, L. Vibrotactile Emotions on a Mobile Phone. In *IEEE Int. Conf. Signal Image Technology and Internet Based Systems (SITIS 2008), Bali-Indonesia* (2008).

[8] Réhman, S., Liu, L., and Li, H. Lipless Tracking and Emotion Estimation. In *IEEE 3rd Int. Conf. on Signal-Image Technology & Internet-based Systems, Shanghai , China* (2007).

[9] Rohs, M. Real-world Interaction with camera-phones. In *2nd Int. Sym Ubi. Comp. Sys. (UCS 2004),Tokyo, Japan* (2004).

[10] Scott, J., Dearman, D., Yatani, K., and Truong, K. Sensing Foot Gestures from the Pocket. In *23rd ACM Sym. User Interface Software and Technology (UIST '10), New York, USA* (2010).

[11] Stichling, D., and Kleinjohann, B. Edge Vectorization for Embedded Real-time Systems using the CV-SDF Model. In *16th Int. Conf. Vision Interfaces (VI 2003), Halifax, Canada* (2003).

[12] Valkov, D., Steinicke, F., Bruder, G., and Hinrichs, K. H. Traveling in 3D Virtual Environments with Foot Gestures and a Multi-Touch enabled WIM. In *Virtual Reality International Conference (VRIC 2010)* (April 2010), 171–180.

[13] Wagner, D., Reitmayr, G., Mulloni, A., Drummond, T., and Schmalstieg, D. Pose Tracking from Natural Features on Mobile Phones. In *7th IEEE Int. Sym. on Mixed and Augmented Reality, Cambridge, UK* (2008).

[14] Weiss, S. *Handheld Usability*. John Wiley & Sons Ltd, 2002.

[15] Yousefi, S., Kondorin, F., and Li, H. 3D Gestural Interaction for Tereoscopic Visualization on Mobile Devices. In *14th Int. Conf. Comp Analysis of Images and Patterns (CAIP 2011), Seville, Spain* (2011).