CAMERA BASED CROSS DEVICES MANIPULATING WITH AUGMENTED REALITY

Teng Li[1], *Wei Hu*[2], *Jun Wu*[3], *and Jinfeng Bai*[4]

1 College of Electrical Engineering and Automation, Auhui University
2 Media Research Lab, Huawei Corp., Beijing
3 Institute of Software Application Technology, Guangzhou & Chinese Academy of Sciences
4 Institute of Automation, Chinese Academy of Sciences
tenglwy@gmail.com, wei.hu@msn.com, wujun@gz.iscas.ac.cn, jinfeng.bai@ia.ac.cn

ABSTRACT

Heterogeneous devices with large displays have become pervasive in people lives. While conveying public information, most of them are lack of support for human interaction. This paper proposes a cross devices interaction system for users to conveniently manipulate the content on public displays through personal mobile phones with camera. Users can acquire, insert, and act on elements of the displayed content with simple operations on mobile phones. Cross device elements targeting is done with an efficient visual based object matching and tracking algorithm when users capture the displayed contents as a live video. Augmented information for the elements is shown on the mobile phone screen. The content in the large displays can then be operated by clicking and dragging, etc., on the mobile phone. The proposed system can be applied to interactive advertisements, multi-user games, and conference discussions. An evaluation consisting of a series of experiments demonstrated the usability and efficiency of the proposed system.

Index Terms— Large display, mobile phone, object matching and tracking, augmented reality.

1. INTRODUCTION

Large displays such as light-emitting diode (LED) monitors and projection screens are used widely in public places such as metros, conference rooms, commercial centers, through which various contents including commercials, presentations, and so on are conveyed from providers to users. In most cases, common users can only receive the information passively without a convenient avenue to interact with them. However, people often have the willing to interact with the displayed content when seeing interesting objects. For example, lecture attendees may want to download some graphs from the projected materials to their mobile phones, and share their personal data on large screens during seminar; Customers like to obtain coupons associated with their interested products on advertising displays and share their own experiences as well.

As the mobile phone with camera has become an indispensable device for each person, it is also a tool for users to acquire, store, and share information. Though they are capable of taking high-quality photos of the displayed things for conserving and analysis, the original elements such as pictures and coupons cannot be obtained directly. Touch screens, which are widely used as interactive interfaces, usually do not support the cross devices content transfer. Neither can they be simultaneously reachable for many users.

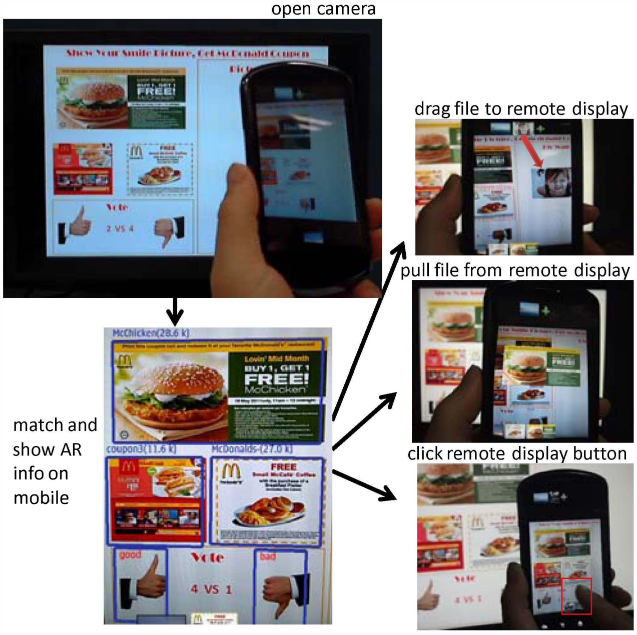
open camera
drag file to remote display
pull file from remote display
click remote display button
match and show AR info on mobile
McChicken(28.6 k)
coupon3(11.6 k)
McDonalds-(27.0 k)
good
Vote
bad
4 VS 1

Fig. 1. A user start capturing content shown on the large display in realtime with a camera equipped phone. Our system targets the displayed elements on phone screen regions via computer vision technique, and corresponding augmented information is pushed to phone screen. Users can then acquire interesting ones to the phone, insert images or texts into specific area, and act with objects by simple operations on phone screen.

In this paper, a novel cross devices system is proposed for users to have thorough interactions with remote public displays conveniently. As shown in Figure 1, the proposed system lets users have such below manipulations with elements in the displayed content.

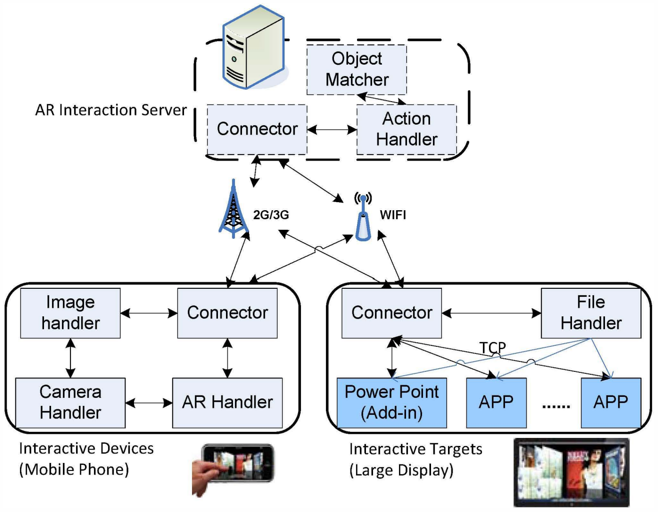

Fig. 2. The architecture of the proposed system: arrow lines represent communicating relations between components.

- *Acquire*: users can transfer interesting elements from the large displays to personal mobile phones in original format, such as excel files inside a presentation.
- *Insert*: users can insert files in their mobile phones into the displayed content, such as pictures and texts.
- *Act*: users can act on the displayed elements to generate response on the large display, and the corresponding content in the large displays can be modified as a results.

All the interactions are done via a camera-equipped mobile phone without any other additional device. User operations on phone screen are as simple as clicking, dragging, and so on. Augmented information for elements is shown on the corresponding phone screen regions when user captures the displayed content as a live video, which gives hints for users to do interaction. The real-time cross devices targeting between content elements on the large display and phone screen regions is built based on the local visual feature based object matching and tracking algorithm. The in-file operations are realized via Microsoft Component Object Model (COM).

The proposed system has the following features that make itself a useful complement to existing cross devices interaction tools. Firstly, it provides a more complete framework for interaction use than ever before, with augmented reality (AR), user operations of acquiring, inserting, and acting. The second, operations and communications are realized in a live video capturing manner, which requires more efficient object matching than taking a picture, and it is realized by the real-time local feature based object matching and tracking here. Finally, users can interact directly with the content inside files, which has not been realized before. Furthermore, the proposed interaction system supports multiple users simultaneously.

In the rest of the paper, after briefly review the related work, we give a detailed illustration of the proposed system, as well as the design and implementation issues behind our technique. The evaluation results and our implemented prototype are introduced then, and finally we draw the conclusions.

2. RELATED WORK

Mobile interaction with other objects based on the recognition of visual code or visual markers has been widely applied. Rohs and Gfeller [1] present to use mobile phone cameras as sensors for the recognition of two-dimensional markers. Want et al. [2] describes using radio frequency identification (RFID) tags to link physical objects to network services. Android phone users can install an application by scanning a quick response (QR) code.

In recent years, there has been an increasing amount of literature on visual pattern matching based cross devices interaction [3], [4]. Shoot & Copy allows identifying the icons on the large display based on the photographic representation and transferring the corresponding content to the mobile phone [5]. Boring et al. [6] computes the homography by extracting the boundaries of rectangular to link the large display and the mobile phone, based on which interactions are taken. However, the global boundary based object matching method they adopted is not robust enough for some cases such as partial occlusion and similar shapes. Deep Shot [7] provides a framework that enables an arbitrary application to migrate not only its content but also its runtime states across devices. These works neither support the manipulations to the content file's elements, nor provide act & response interactions to the remote display.

Augmented reality has been used in interaction ever since Kato et al. [8], which utilized markers to extract the accurate position and orientation of a video camera. The recognition of markers is an established technology for the identification of augmented objects and the interaction with them. The step to mobile devices has been made in recent years [9]. However, cross devices interaction has not been addressed here.

3. THE PROPOSED INTERACTION SYSTEM

The proposed interaction framework as shown in Figure 2 consists of three main components: interactive target e.g. the large advertisement screen, meeting room projections, which is used for presenting information to the public; interactive device e.g. the camera-equipped mobile phone; and interaction server, which crawls the information from and send response to the interactive target; the interaction server plays the role of coordinating the communication between the interactive device and the interactive target.

The interactive target contains three components: the running applications (App), the file handler and the connector. The running App displays the information to the public. For example, a presentation slide is projected to the screen in a seminar and the running App is PowerPoint in this case. The file handler tracks and parses screen-shot of the running App, and responses to the actions on the running file requested by the server. The connector is responsible for communication and data exchanging between the server and the interactive target.

The interactive device is to obtain the interesting information presented on large screen in detail from the interactive target and send the feedback. The camera handler controls the mobile phone to acquire the visual data as an image sequence. The image handler then encodes the captured image into Joint Photographic Experts Group (JPEG) format for data communication. The message containing the information of captured region and requested object of the mobile phone side is sent to the server via the connector. Once the connector receives response from the server, the connector notifies the AR handler to display the retrieved information on the phone screen. The AR handler also collects user operations and sends them to the server via the connector.

The interaction server takes the task of processing requests sent from both the mobile phone side and target side. The server receives the screen-shot from the interactive target and the captured image from the mobile phone. Objects in these two pictures are matched using computer vision technique [10]. A coordinate switcher is then used to map the object captured by the phone camera and the object displayed on the presentation screen. Once the displayed objects are targeted to the phone screen successfully, the mobile phone and target can communicate via the server. For example, the server acquires the information that the mobile phone requests an object from the target, and sends the object to the mobile phone. The mobile phone can also send its feedback to the server and then the server will pass the information to target so that the feedback will be displayed instantly on the presentation screen. The server component is independent to the target and the mobile phone. The framework can be easily extended to multi-user and multi-target cases.

The key technique for the system usability is to match visual targets based on the captured images sequence fluently and make the user interaction smoothly. To this, the communication and image processing modules are carefully designed. The file handler is to realize the in-content manipulations on the files. An efficient local feature based object matching and tracking algorithm is adopted here. These technical components are detailed in the following.

3.1. The connector

The socket protocol based on Transmission Control Protocol (TCP), which is widely adopted in data transmission tasks, is used here for communication between the server and the devices. There are four types of messages to be transferred: image region matching information, content acquired from the display to the mobile phone, data to be inserted from the mobile phone to the large display, and information of user actions on the mobile phone. All messages are transmitted in packages with a unified format as [command] | [detailInfo]. The detailed package format for these types of transmitted data is given as follows:

- Match: match | [previewImageData] | [locationInfo]. The [preview ImageData] is generated from the camera preview image which is converted to hex string for transferring. [locationInfo] could includes global positioning system (GPS) information, i.e., latitude, longitude, general packet radio service (GRPS) information, i.e., Cellid, location area code (LAC) or wireless fidelity (WIFI) information, i.e., service set identifier (SSID). All information could help to identify which display should be connected.
- Insert: insert | [coordinateX], [coordinateY] | [objectType] | [objectName] | [objectData], where the coordinate represents the positions that the user drags local files from mobile phone to remote display through camera preview. The [objectType], e.g. IMAGE, TEXT, orMP3, is to tell the file handler in interactive target the content type to be inserted.
- Acquire: acquire | [objectName]. After the user clicks an object in phone camera preview, the targeted content element in remote display is mapped. Then the acquired information is sent to the server.
- Act: act | [objectName] | [actInfo] | [eventInfo], the detail description of action.

To efficiently utilize the network resource, each device has a shared socket connection with exterior devices. Asynchronous socket connection is adopted to speed up the communication. Therefore, devices in the framework could connect to each other effectively and multiple users can be supported. Figure 3 shows the flowchart of matching and acquiring action procedure.

3.2. Object matching and tracking

Once the user captures a region on the target screen, the corresponding augmented information is shown on phone screen in addition to the displayed content. The accurate objects mapping between the phone screen and the targeted screen-shot is required, which is important for identifying which content element the user acts on. Figure 3 gives the whole procedure of objects matching and tracking.

Since there are usually translation, rotation and scale variations between the phone screen and the large display screen, as well as the partial occlusion. The local visual feature based matching algorithms such as scale-invariant feature transform (SIFT) [11] and Speeded-Up Robust Features (SURF) [10], have proven their robustness to these variations in object matching comparing to the global feature based matching algorithms [12]. The SURF matching is adopted here for its better computational efficiency than SIFT, as well as good matching accuracy. Matching between the captured region and the screen-shot is done on the server.

As shown on Figure 4, usually a set of SURF feature points can be detected from each image and each of them is represented by a feature descriptor vector, which is calculated with the SURF algorithm, using the methodology proposed in [10]. To find the feature correspondences between images, the nearest neighbor in the other image for each local feature in term of dot product distance between SURF descriptors is firstly found, and obtain several candidate matches. Then some false matches are rejected according to the log likelihood ratio if it is above a *distRatio* threshold of 0.8. The ratio is defined as the potentially best matched descriptor (val_1) to its next best matched descriptor (val_2) as illustrated in the following equation:

$$Match = \begin{cases} 1, if \ \frac{val_1}{val_2} < distRatio \\ 0, \ otherwise \end{cases} . \qquad (1)$$

Figure 4 shows an example of the feature matches obtained after this step. We can see that some feature matches returned by this process are still incorrect. In order to verify matches between two different images captured at different viewpoints, the homography is computed to reduce the effect of incorrect matches, by the random sample consensus (RANSAC) approach [13].

To compute the local feature matching continuously for each frame the mobile phone captured is time-consuming and would impede fluent user interaction. After the corresponding objects are targeted, the temporal continuity of captured frames is utilized here and the Mean Shift tracking algorithm is applied [14] to track the objects for better computational efficiency.

3.3. Filer handler

The file handler takes charge of the file operation related tasks including:

- Obtaining current file information, such as state changing, information of current shown elements.
- Modifying content according to the user operations, such as inserting a picture or some text to a specific area.
- Generating response when the user acts on some elements from the mobile phone.

An add-in for interacting with Office series files is implemented. The supported file formats include slides, documents, and etc., which are commonly used in large displays, based on the COM interface provided by Microsoft in C#. Status monitoring and content manipulation for the files can be done by the add-in.

4. TECHNICAL EVALUATION AND USER STUDY

To analyze and evaluate the proposed interaction framework, a prototype system was developed. A desktop has been used as the interaction server and master of the big display. A 22-inch LED monitor with 1440*900 resolution is used to simulate the interactive target. In addition, the mobile part of the software was deployed on a smart phone running Android 2.6.32, which serves as the interactive device. The desktop and the mobile phone are connected to each other by a WIFI route. This environment is easy to set up by common users, based on which both objective technical evaluation and subjective user study for the proposed system were conducted. The detailed evaluation results are given in the follow.

4.1. Technical evaluation

The transmitted camera captured data from the mobile phone to the server is important for system usability. The accurate object matching relies on enough data information whereas transmitting large data would impede communication efficiency. Two related issues need to be considered here: what kind of data to transfer, original images or extracted visual feature data, and what image resolution to use. These factors influence the required transmitted data size from the mobile phone to the server. In the following two experiments are conducted to study the matching accuracy and communication speed regarding different transmission choices and image resolutions.

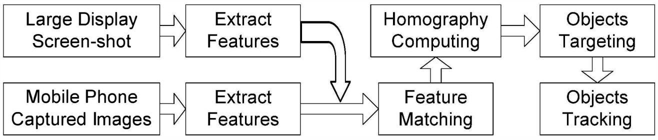

Fig. 3. The flowchart of objects matching and tracking procedure between large display screen and mobile phone captured images.

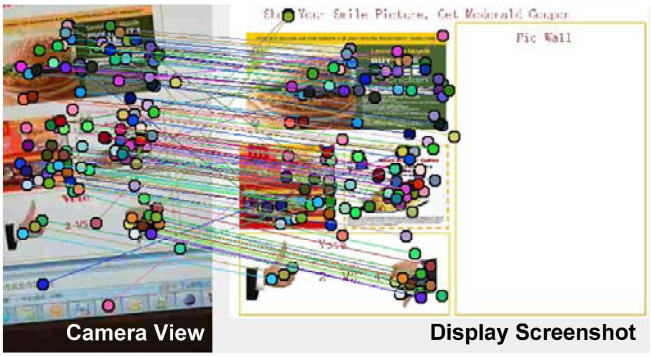

Fig. 4. The image matching algorithm finds correspondences between the phone captured picture (left) and the remote display screenshot (right) before computing homography.

Experiment 1: To transmit images or feature data

When the remote display showed an advertisement in full-screen mode, pictures of the display screen are taken in different distances (20cm, 30cm, 40cm, 50cm, 60cm, 70cm, 80cm) and different viewpoints (20° to left, center, 20° to right, 20° to below) with the mobile phone. For each case two pictures were taken, as a result 56 pictures in total were obtained for testing. The picture resolution is 360*600. The success matching rates and feature data sizes are obtained with different feature numbers when varying the parameter of SURF algorithm.

As Figure 5 (a) shows, the matching rate is up to 90% when extracting 329 interest points by SURF. The 329 interest points data, including coordinates and visual description, occupies approximately 180K bytes in byte array format, which is much larger than the average picture size i.e. 14K bytes. Figure 5 (b) shows the detailed comparison result. When using 50 interest points only, the visual feature data size is similar to that of the original picture, but the matching rate drops down to 30%, which is unacceptable.

Therefore, original picture captured by mobile phone camera is a better choice to transmit than visual feature data. Furthermore, to use the original images is convenient for the server, which is computationally powerful, to apply better image processing and matching algorithms.

Experiment 2: Image resolution for performance

This experiment studies image file size, interest point number, matching time cost and matching rate with different image resolutions, to find a reasonable resolution for visual data transmission and pattern matching. Four different image resolutions are tested here: 120*200, 240*400, 360*600, and 480*800. All smaller images are scale-shrink from the 480*800 images.

Figure 6 (a) shows that the image file size increases as the resolution increases. The image file size is about 22K bytes when the image resolution is 480*800, so that if three images are transmitted per second, the network bandwidth needed is 66K bytes per second at least. When the transmission image resolution is in 240*400, only 21K bytes per second of the network bandwidth is required, with the match rate drops to around 65%. As shown in Figure 6 (b) and (d), the number of interest points and the matching rate also increase as the image resolution increases using the same parameters setting in SURF matching. Figure 6 (c) shows the relation between the image resolution and the matching time.

From the above test results, we can see that 360*600 is the best resolution choice considering the balance between network bandwidth resource and visual matching rate, when the matching rate is 85% and the required network bandwidth is 39K bytes per second. The bigger resolution could be used in the high-speed broadband network such as WiFi, 3G network, and vice versa.

Since the test pictures were taken from different distances to the large display, i.e., ranging from 20cm to 80cm, and different viewpoints as explained in the former experiment, the average matching rate is not good enough due to the pictures taken from long distance and angled viewpoint. Our experiments shows that a proper distance and viewpoint setting can yield high success matching rate.

4.2. The implemented prototype

As aforementioned, the proposed system allows users to interact with the displayed content in a remote distance, including those that would be not easy for multiple users to reach simultaneously, such as projection wall, and advertisement LED. Users aim the device with a mobile phone camera and the displayed content will be played on the phone as a live video. Augmented information marks associated with content elements are shown together in real time.

In order to allow mobile use, the proposed system continuously computes the mapping between phone screen regions and the displayed content by visual patterns matching, which is done using the phone camera. It identifies displayed elements on phone screen to transmit the user interaction on the mobile device into the target remote display, thus provides a useful avenue for content providers and common users. Two typical scenarios that can be accomplished are illustrated in the below.

- *Interactive advertisement*: This scenario can be useful for both advertiser and customers. Currently, there are many advertisements playing on the public displays everyday. When people watch them, they may want to download the coupons for interested products to their mobile phones, and give their feedbacks. To these aims, they usually first recognize the product names, then open the browser on the phone and search the products, and finally reach a commercial website. In this scenario, the proposed system provides a much more convenient way for users to carry on these operations. Moreover, the advertisers can convey more information for products through augmented reality, and collect user data for better propagation.
- *Interactive presentation*: Users can insert photos and texts in mobile phones into the displayed content, as well as transfer some interesting elements of the presentation to personal devices with simple shooting the display and touching phone screen operation. It is useful to share information between mobile phones owned by different people, and therefore facilitate the communications in meetings or seminars.

4.3. User study

An initial user study has been conducted to observe if people would use our prototype. Our main goal is to find out if the system is attractive by users and merchants for interactive advertisements. We were also interested in the acceptance of users according to our scenarios.

Figure 7 gives the investigation results. There were 50 participants in this study which consists of 30 common users and 20 merchants. We collected comments from people in the street as common users. Their ages range from 15 to 37 with an average of 24.5. Two users expressed that they never paid attention to the information on big display outside. Most of them showed interests in the novel interaction with big display especially in downloading coupons conveniently from the advertisements on big screens. The interviewed merchants are from a variety of professions, such as shop owners, restaurant owners, supermarket managers, and online dealers. Their ages range from 24 to 53 with an average of 32. All of them have the experience to advertise on big displays in subway stations, streets or shopping malls, and they cared about how to attract more people to their products. Most merchants in this investigation were glad to try this system, though some of them worried about the extra cost for the new type of advertisements.

5. CONCLUSIONS

A novel cross device interaction system is proposed, and it allows a set of user operations on content shown on a distant display through a mobile phone equipped with a camera. The interaction is performed on a live captured video with augmented information, and the users can operate directly on content elements shown on the display the supported interactions and usability have not been realized in previous work. The presented framework supports multiple users and can be extended to applications with multiple interactive targets and multiple interactive devices. It therefore provides a useful tool for both content providers and common users in scenarios such as interactive advertising, conference presenting and interactive games. To allow the accurate and real-time matching between phone screen regions and targeted content elements, a fast local feature based image matching method is used and extensive studies on the usability have been done in experiments. Both objective and subjective evaluations show the proposed system is promising.

6. REFERENCES

[1] M. Rohs, B. Gfeller, “Using camera-equipped mobile phones for interacting with real-world objects”, In *Proceedings of Advances in Pervasive Computing*, 2004.

[2] R. Want, K. Fishkin, and A. Gujar, “Bridging physical and virtual worlds with electronic tags”, In *Proceedings of the international conference on human factors in computing systems (CHI)*, 1999.

[3] C. Liao, Q. Liu, B. Liew and L. Wilcox, “Pacer: fine-grained interactive paper via camera-touch hybrid gestures on a cell phone”, In *Proceedings of the 28th international conference on human factors in computing systems(CHI)*, pp. 2441-2450, 2010.

[4] Q. Liu, P. McEvoy, and C.-J. Lai, “Mobile camera supported document redirection”, In *Proceedings of ACM Multimedia*, 2006.

[5] S. Boring, M. Altendorfer, G. Broll, O. Hilliges, and A. Butz, “Shoot & Copy: phonecam-based information transfer from public displays onto mobile phones”, In *Proceedings of the 4th international conference on mobile technology, applications, and systems (Mobility)*, pp. 24-31, 2007.

[6] S. Boring, D. Baur, A. Butz, S. Gustafson and P. Baudisch., “Touch Projector: mobile interaction through video”, In *Proceedings of the 28th international conference on human factors in computing systems(CHI)*, pp. 2287-2296, 2010.

[7] T.-H. Chang, Y. Li, “Deep Shot: a framework for migrating tasks across devices using mobile phone cameras”, In *Communications of the ACM*, 1993.

[8] H. Kato, M. Billinghurst, “Marker tracking and HMD calibration for a video-based augmented reality conferencing system”, In *Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR)*, 1999.

[9] J, Reid,R. Hull, K. Cater and C. Fleuriot, “Magic moments in situated media scapes”, In *Proceedings of CHI International Conference on Advances in Computer Entertainment Technology*, 2005.

[10] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: speeded up robust features”, *Computer Vision and Image Understanding*, Elsevier, 2008.

[11] Lowe, D. G., “Distinctive image features from scale invariant keypoints”, *International Journal of Computer Vision*, 60:91–110, 2004.

[12] K. Mikolajczyk and C. Schmid, “Performance evaluation of local descriptors” *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005.

[13] Fischler and Bolles, “Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography”, *Commun. Assoc. Comp. Mach.*, 24:381–395, 1981.

[14] D. Comaniciu and P. Meer, “Mean Shift: a robust approach toward feature space analysis”, *IEEE Trans Pattern Anal. Machine Intell.*, Vol. 24, (5), pp. 603-619, 2002.

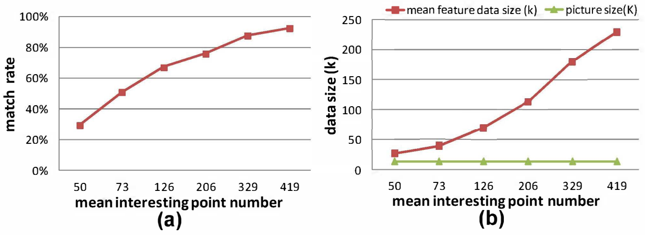

Fig. 5. Comparisons of matching rates (a) and feature data sizes (b) using different interest point numbers when varying the parameter of SURF feature extraction. The original picture size is 14k bytes.

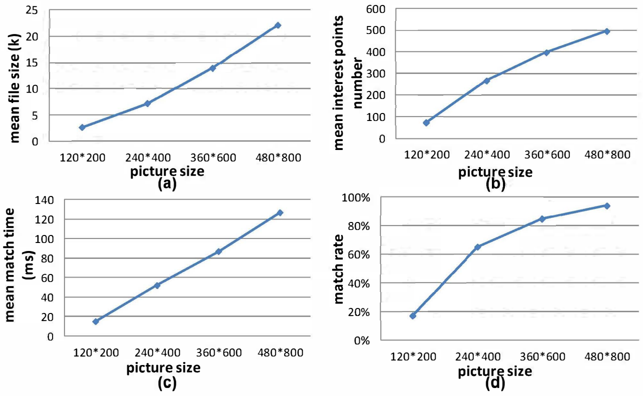

Fig. 6. Comparisons of data sizes (a), detected feature numbers (b), matching times (c) and matching rates (d) using different picture resolutions. The average result values of 224 pictures captured in different distances and viewpoints are showed, with a fixed parameter for SURF feature extraction. A proper picture size for transferring can be chosen accordingly.

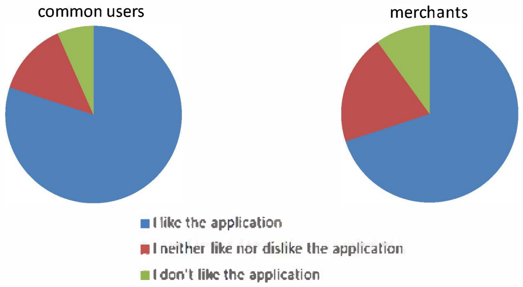

Fig. 7. The subjective study result statistics: the left pie chart is for common users and the right one is for content providers such as merchants.