2012 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 7-12, 2012. Vilamoura, Algarve, Portugal

Embedding Imperceptible Codes into Video Projection and Applications in Robotics

Jingwen Dai and Ronald Chung

Abstract— **We describe a method of embedding imperceptible codes, that are for structured light sensing (SLS), into regular display in video projection. To resolve the conflict between imperceptibility of the embedded codes and feasibility and robustness of SLS code retrieval, we introduce antinoise schemes to the coding and decoding stages simultaneously. On one hand, specially designed primitive shapes and large Hamming distance are employed in the spatial SLS coding to enhance the error-tolerance capability. On the other hand, pre-trained primitive shape detectors are applied to detect and identify the embedded codes in the decoding stage – a task difficult to achieve by segmentation that is adopted in classical structured light methods, when weakly embedded information is inevitably interfered by noise. Extensive experiments show that the proposed method is effective and accurate in code retrieval, even with the prerequisite of incurring minimum disturbance to arbitrary video projection. Some potential applications to a robotic system are also demonstrated.**

I. Introduction and Motivation

The improving performance and declining price of digital video projectors make it possible to use them prevalently. Being able to generate arbitrarily large display is a feature of projectors that makes them exceedingly attractive, especially in applications that demand portability.

On the other hand, the adoption of structured light illumination has been proven to be an effective and accurate means for 3D information perception [5]. Recently, the availability of pico projectors with average dimensions of $4 \times 2 \times 1$ inches has widely extended the application domain of structured light system. There are already pocket DCs, DVs and cellular phones in the consumable market that have both projector and camera built-in, making it possible to implement structured light system in hand-held consumer electronic products.

In other words, projector accompanied by camera has the potential of being a device for both display and sensing, i.e., for both input and output in human-computer interface, making it a possible device to replace traditional LCD panel, keyboard, and touch-sensitive screen altogether in computing, at the cost of only diminished size and weight. Projector has the potential of making a breakthrough of dramatically downsizing portable computing without sacrificing display size.

For these reasons projector-camera (ProCam) system has been actively researched in the last few years. Many research groups apply projectors in unconventional ways to develop new and innovative information displays that go beyond simple screen presentations [2].

Some researchers designed structured light system in the non-visible spectrum [4]. That way the media for regular projection and structure light sensing can be made separate. However, additional hardware could be reduced and device size could be diminished if structured light and regular projection can be achieved through the same projector. This leads to the concept of Imperceptible Structured Light (ISL). ISL modulates the projected display either spatially or temporally to embed code patterns for structured light sensing. In principle, due to limitation of human visual perception, the embedded code patterns can be made undetectable to the user, but cameras synchronized to the modulation are able to reconstruct the embedded codes for structured light sensing.

There is however challenge in embedding codes into regular projection. While the codes should be made as undetectable as possible to the user, they have to be decodable to the camera for the purpose of structured light sensing. On top of the dilemma, there is the inevitable fact that the displayed signals are generally corrupted by substantial noise that arises from the nonlinearity of the projector, the sensing defects of the camera, and the variation of the ambient illumination. The objective of this work is to deal with the dilemma.

This article describes a novel method of embedding imperceptible structured codes into arbitrarily intended projection. Through precise projector-camera synchronization, structured codes consisting of three primitive shapes are embedded into the projection, in a way that is imperceptible to viewers but extractable from the "difference image" between successive images captured by a camera. To make the decoding process more robust against noise, we do not extract the codes by region segmentation in the image domain. Instead we employ specially trained classifiers to detect and identify the codes. To enhance the error tolerance further, specially designed primitive shapes and large Hamming distance are adopted in the spatial coding. Even with some bits of the codewords missed or wrongly coded, the correct correspondence could still be derived correctly. Since the method is efficient and robust in 3D sensing, it has great potentiality to integrated into robot system for many applications.

The remainder of this paper is structured as follows. In Section II, related works on imperceptible structured light sensing are briefly reviewed. The principle of embedding imperceptible codes along with robust coding and an antinoise decoding mechanism are described in Section III. In Section IV, system setup and experimental results are shown. Some

Jingwen Dai and Ronald Chung are with Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong {jwdai,rchung}@mae.cuhk.edu.hk

potential applications in robotic system are demonstrated in Section V. Conclusion is offered in Section VI.

II. Related Works

A first proof of concept for embedding invisible structured light patterns into DLP(Digital Light Processing) projections was introduced in the "Office of the Future" project [6]. In this work, binary codes are embedded by projecting temporally alternating code images and their complements. Provided that the frequency of projection reaches the *flicker fusion threshold* ($\geq 75Hz$), the pattern and the inverse pattern are visually integrated over time in human perception, and the illumination has the appearance of a flat field ("white" light) to humans. However, the concept of embedding structured light into DLP projections was achieved with significant modification effort on the projection hardware and firmware, including removal of the color wheel and reprogramming of the controller. The resulting images were also in greyscale only. The implementation of such a setting was impossible without mastering and full access to the projection hardware.

Cotting et. al. introduced a coding scheme [7] that synchronizes a camera to a specific time slot of a DLP micromirror flipping sequence in which imperceptible binary patterns are embedded. However, not all mirror states are available for all possible intensities, and the additional hardware, DVI repeater with tapped vertical sync signal, is not an off-the-shelf instrument.

However, with the development of digital projection technology, some so-called 3D compatible DLP projectors with fresh rate of $120Hz$ or higher emerged recently. This makes it possible to implement imperceptible structured light without any hardware modification or extra assisting hardware. Many researcher began to study how to determine the embedded intensity properly to guarantee the code imperceptibility.

In [1], subjective evaluation results and their statistical analysis on the visual perceptibility of embedded codes in different ways were reported. The factors affecting code visibility are also concluded. Park et al. [12] presented a technology for adaptively adjusting the intensity of the embedded code with the goal of minimizing its visibility. It was regionally adapted depending on the spatial variation of neighboring pixels and their color distribution in the YIQ color space. The final code intensity was then weighted by the estimated local spatial variation. Since two manually defined parameters adjusted the overall strength of the integrated code, the system was not able to automatically calculate an optimized intensity. Grundhofer et al. [13] proposed a method considering the capabilities and limitations of human visual perception for embedding codes. It estimated the Just Noticeable Differences (JND) based on the human contrast sensitivity function and adapted the code intensity on the fly through regional properties of the projected image and code, such as luminance and spatial frequencies. The shortage of this method was that some parameters needed be pre-measured using some optical devices (e.g. photometer), which were not accessible to nonprofessional users.

To the best of our knowledge, up to now, seldom works focus on the decoding method in imperceptible code embedding configuration, especially, when huge external noises exist.

III. Approach

A. Principle of Embedding Imperceptible Codes

The fundamental principle behind imperceptible structured code embedding is the temporal integration achieved by projecting each image twice at high frequency: a first image containing actual code information (e.g., by adding or subtracting a certain amount (Δ) to or from the pixels of the original image, depending upon the code) and a second image that compensates for the distortion in the first image. The vital aspects of ISL sensing are code embedding and projector-camera synchronization.

Since general projection is in color, it is possible to embed color code through three different channels theoretically. However, to enhance code robustness toward noise, we use binary code and embed it into all three color channels simultaneously. Let B, O, I and I' be the binary code image, the original image, the projected image, and the complementary image respectively. Then the projected image and complementary image could be formulated as

$$I_i(x,y) = O_i(x,y) + P(x,y), \quad (1)$$

$$I'_i(x,y) = O_i(x,y) - P(x,y), \quad (2)$$

$$P(x,y) = \begin{cases} \Delta, & when \quad B(x,y) = 1; \\ 0, & when \quad B(x,y) = 0. \end{cases} \quad (3)$$

where i = {R,G,B} indicates red, green and blue channels, Δ is the embedded intensity.

To avoid intensity saturation at lower and higher intensity levels when adding or subtracting Δ, the original image needs to have the intensity range in each color channel compressed to between Δ to $255 - \Delta$. Since the embedded intensity required in the coding is small enough, the visual degradation due to contrast reduction is negligible.

The degree of imperceptibility thus depends upon the embedded intensity. A larger intensity ensures that the code be more tolerant toward noise and more readable in the image of the projection, whilst a smaller intensity makes the embedded codes more invisible. In our design, code imperceptibility has higher priority, and thus embedded intensity is set to a very small value.

In order to achieve imperceptible structured light projection, the frequency of projection must exceed the flicker fusion threshold, which is $75Hz$ for most of the people. Here we take one projection-capture cycle as an example to explain the strategy of projector-camera synchronization, which is illustrated in Fig. 1. Firstly, we ensure that the projector projects an image every $10ms$, i.e., at $100Hz$. As shown in Fig. 1, along the time axis, the projected image I and the complementary image I' are projected at the time instants $0ms$, $10ms$ respectively. With a refresh rate of the camera at about 100 frames per second, the camera captures the image C and C' at $5ms$ and $15ms$, shortly after the

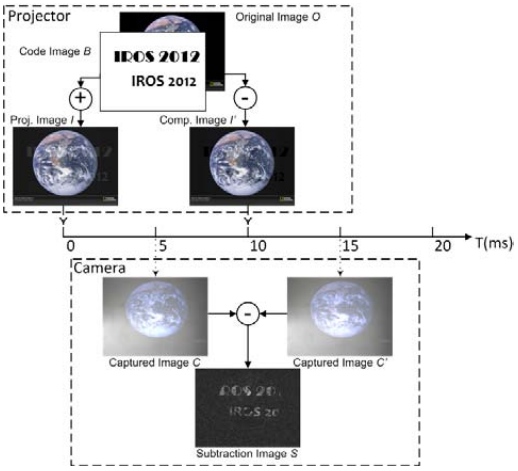


Fig. 1. Projector-camera synchronization and basic principle for embedding and extracting imperceptible codes

projector projects the projected image and complementary image to the scene. At 20*ms* a new projection-capture cycle will resume. With the aforementioned projection-capture strategy, the system could capture 50 image pairs per second.

The embedded codes could be internally and simply extracted from the "subtraction image"[1] between consecutively captured images as

$$S(x,y) = \max_i [C_i(x,y) - C'_i(x,y)], \quad i = \{R,G,B\}. \quad (4)$$

Ideally, the subtraction image should be a binary image that has maximum value of 2Δ and minimum value of 0. However, the subtraction image in reality is generally disturbed by large external noises, as shown in Fig. 1. Since the embedded intensity is always small ($\Delta \leq 10$), the subtraction image has low signal-to-noise ratio. It is generally nontrivial to retrieve the embedded codes. In the rest of this section, we describe how robust coding and noise-tolerant decoding approaches can help tackle the issue.

B. Design of Embedded Pattern

The strategy of encoding in general structured light methods could be classified into two categories [5]: time multiplexing and spatial multiplexing. The former one can achieve denser data samples with higher accuracy, but at the expense of requiring multiple illuminations and image captures over time, which is not suitable for imperceptible code embedding [1] and dynamic scenes. In contrast, the latter one labels each pattern position by the appearance profile of the neighboring positions. The appearance profile can be about various gray levels, colors, or geometric primitives, and the coding methods include De-Bruijn sequences, pseudorandom arrays, and M-arrays [5]. The spacial coding scheme has the advantage that 3D determination could be achieved with a single pattern.

Considering the constraints of imperceptible code embedding, we employ the spacial multiplex scheme to design our pattern. Due to the choice of using binary code for robust code embedding, the symbols cannot be coded with different colors, so we use an alphabet set comprising three different geometrical primitives: cross, sandglass, and rhombus, as shown in Fig. 2. There are three advantages of this configuration. First, all the shapes own a natural center point, which simplifies the shape identification process in the decoding stage. Then, there are sufficient variations between different shapes; even with large disturbance from noise on the shapes, the decoding method could distinguish them. Moreover, the directional information carried by the cross shape could rectify the observation window during the step of neighborhood detection without enforcing any other constraints.

Fig. 2. The primitive shapes: cross, sandglass and rhombus

In the decoding stage, the centroid of each detected primitive would be considered as the feature point position, and the 9-bit codeword associated to each feature point is composed of the elements in the 3×3 window centered on it. In traditional structured light methods, the uniqueness of the codeword is usually assured by M-arrays (perfect maps), which are random arrays of dimensions $r \times v$ in which a submatrix of dimensions $n \times m$ appears only once in the whole pattern [8]. The M-arrays give a total of $rv = 2^{nm} - 1$ unique sub-matrices in the pattern and a window property of $n \times m$. However, the Hamming distance between the codewords is 1, which is generally too small for our code embedding scenario in which the codeword retrieval errors could be large due to noise. In our system, we generate a matrix of dimensions 27×29 using the method proposed by Albitar [9], in which 95.97% of the codewords have a Hamming distance higher than 3 and the average Hamming distance is $\bar{H} = 6.0084$, so that even some bits in the codeword are missed or incorrectly coded, the codeword is still distinguishable. On the basis of this matrix, the binary code image composed from the primitive shapes appears like the one illustrated in Fig. 3, in which the size of each primitive shape is a collection of 11×11 pixels while the interval between each shape is 11 pixels. The total number of feature points is 783.

C. Primitive Shape Identification and Decoding

In the decoding stage, the existence of intense noises (from projector projection, camera sensing, ambient illumination and object surface reflection influence) makes it impossible to segment the primitive shape by the integrated use of region segmentation and edge or contour detection as in ordinary structured light methods. Here, we regard the primitive shapes as objects to "identify" and "detect" rather than "segment".

Compared with other object identification or recognition methods, the machine learning approach proposed by P. Viola [10] has been shown to be capable of processing images rapidly with high detection rates for visual object detection.

[1]All the subtraction images in this article are scaled to [0, 255] for illustration purpose.

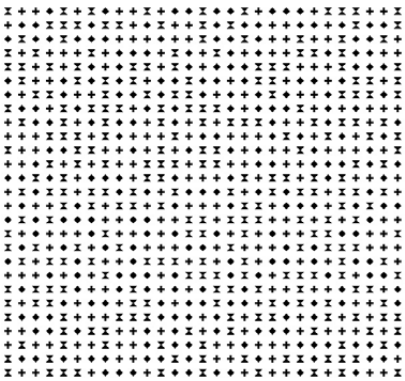


Fig. 3. The embedded binary code image

The approach is adopted here for training detector to identify the three primitive shapes. Below we use cross shape as an example to describe the procedure of detector training.

The performance of training-based detector has a great deal to do with the availability of training samples. Unlike generic objects like human face, body or vehicle, which have a large number of samples in a great many of public databases, we have to collect the specific training samples ourselves in the required configuration. 500 color images with different contents were collected from Internet, and 40 cross shapes were embedded in those images at different positions to generate 500 pairs of projected images and complementary images. By projecting them to a locally smooth textureless surface with orientation variations, 500 subtraction images could be derived from image capture exercises. The sub-images containing cross shapes were then segmented by manual labeling, which were considered as positive training samples. The background images with holes filled by random noise were divided into small patches to generate negative training samples. The training sample preparation process is shown in Fig. 4.

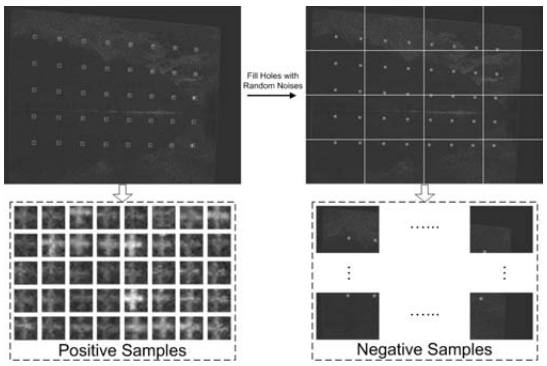


Fig. 4. Training sample preparation

To obtain the optimal performance, the positive samples were resized to 20×20, the extended haar-like features and Gentle Adaboost algorithm were employed, following the suggestion in [11]. Eventually, from over 7000 positive samples and 3000 negative samples, a 16-stage cascade classifier for cross detection was trained. Following the same procedure, the detectors for sandglass and rhombus shapes could be derived as well.

By using the pre-trained primitive shape detectors, the centroid of each primitive, i.e., the position of each feature point, can be determined. Once a feature point is extracted from the image, its codeword can be produced from the associated 3×3 intensity window centered on the feature point. Its corresponding point on the projector image plane is known a priori. This way 3D position on the object surface can be determined via triangulation. The above is the 3D sensing step we use in the system.

IV. Experiments

A. Experimentation Setup

To assess the feasibility of the proposed method for embedding imperceptible codes in regular projection, we conducted experiments on accuracy evaluation.

The projector-camera system we used consisted of a DLP projector (Mitsubishi EX240U projector) of 1024×768 resolution and $120Hz$ refresh rate, and a camera (Adimec OPAL-1000 CCD camera with Myutron FV1520 f15mm lens) of 1024×1024 resolution and $123fps$ frame rate, both being off-the-shelf equipments. The focal length of the camera was fixed at $15mm$, while that of the projector was in the range of $25-31mm$. The ILS was configured for a working distance (the distance from the camera to the mean position of the working area) of about $800mm$.

We first fixed the camera and projector rigidly, and the projector and camera were connected to a desktop computer through VGA and Camera Link interfaces respectively. Then the projector-camera system was calibrated using an LCD monitor as the calibration object; the calibration method, detailed in [3], could derive the intrinsic and extrinsic parameters of the two instruments. Once the experimental system was set up and calibrated, we could conduct further experiments.

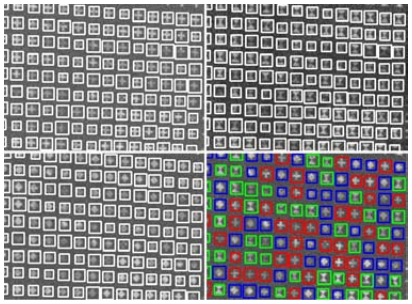
B. Accuracy Evaluation

To assess accuracy, the experimental data with ground-truth were required. Three different primitives and the spatially coded pattern image were embedded into 500 images used for imperceptibility evaluation respectively, with intensity $\Delta = 10$. Then the projected and complementary images were projected successively to a smooth surface, while the camera conducted synchronized capture. The surface was adjusted to different positions and orientations with respect to the camera to involve sufficient shape distortion in the test data. Then the subtraction images embracing embedded codes information were derived for accuracy evaluation. The ground-truth was obtained by manual labeling in the image data captured under binary pattern illumination.

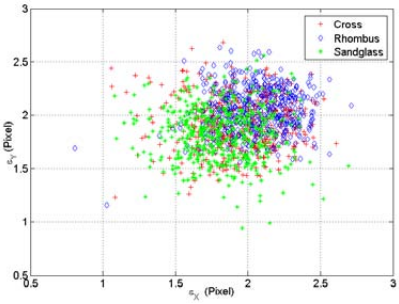
Experimental results in some subtraction images are presented in Fig. 5 (a). The four sub-figures display the cross (top-left), sandglass (top-right), rhombus (bottom-left) shapes, and the spatially coded pattern (bottom-right) respectively. For qualitative evaluation, the detected features are indicated by rectangles, and in bottom-right sub-figure, the

cross,sandglass and rhombus shapes are separately marked by red, green and blue rectangles. The average feature point detection errors along the x-axis and y-axis (as shown in Fig. 5 (b)) are formulated as $\varepsilon_X = \frac{1}{N}\sum_{i=1}^{N}|X_d - X_g|_i$, $\varepsilon_Y = \frac{1}{N}\sum_{i=1}^{N}|Y_d - Y_g|_i$, where N is the total number of embedded shapes, (X_d, Y_d) and (X_g, Y_g) are the detected coordinate and ground-truth respectively. The more detailed quantitative testing results are listed in Table I. Through the proposed method, 91.23% of the embedded feature points could their correspondences found correctly.

By analyzing the missed and false detection cases, we find that the mistakes were mainly caused by large noise that occludes the embedded codes, implying that external noise has the greatest influence on the decoding process.



(a)



(b)

Fig. 5. (a) Some qualitative experiment results. (b) The average detection error of three primitives.

V. Potential Applications in Robot System

For the purpose of illustrating the proposed method's potential applications in robotic system, we mounted a projector and a camera rigidly on special designed frame, and then fixed the frame on a tripod affixed on a mobile robot manufactured by ARRICK Robotics [14], as shown in Fig. 6 (a). Considering the mobility and weight, TI Pico Projector Development Kit [15] and Point Grey Flea3 CCD camera were adopted, as shown in Fig. 6 (b).

A. Sensing Surrounding Environment

For a mobile robot, one of the essential capabilities is to sense the surrounding environment for navigation, obstacle avoidance, object recognition and some other purposes. We assist the visual sensing through a normal grey illumination with invisible codes embedded. By retrieving the embedded

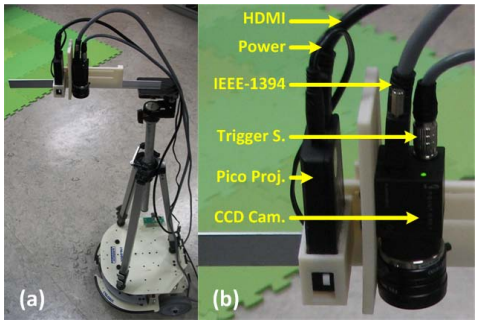


Fig. 6. Integration with mobile robot system

codes, correspondences between projection plane and image plane could be established accurately and efficiently. In Fig. 7 (a) and (c), a green tea can and toy bricks were located in the illumination area of the projector, 3D depth information of certain points on the objects was acquired through simple triangulation in real-time. The surfaces of the objects were rendered in 3D as shown in Fig. 7 (b) and (d). Although the ground truth of the objects was not available, such qualitative examinations showed that the reconstructed surfaces were of reasonable quality.

It is worth pointing out the texture of the object would not effect the 3D sensing results just as the example of green tea can , by reason that in our method the decoding process was conducted in subtraction image, which would weaken the texture influence to a certain extent.

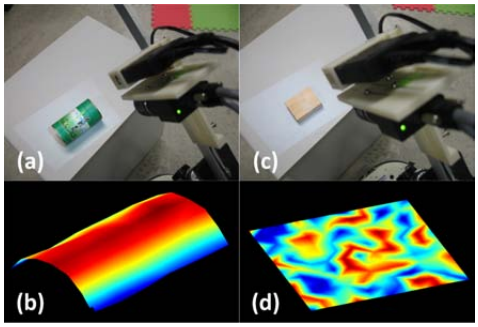


Fig. 7. Some 3D sensing results

B. Natural Human-Robot Interaction

Besides sensing capabilities, the mobile robot should also provide an effective channel for the interaction between users, such as an interface for system configuration or a display panel to show prompt information. In traditional way, an LCD monitor plus mouse-and-keyboard or an LCD touch-screen are attached to the robot, inevitably increasing the weight and size of mobile robot, let alone more energy consumption. Our method enables a common projector to serve the dual role of a display device as well as a 3D sensor with the assistance of camera, providing a platform for more natural user interface schemes. As shown in Fig. 8 (a), a system configuration interface (Fig. 8 (b)) was projected onto a desk surface, a user was operating on the projected desk

	Hits(%)	Missed(%)	False(%)	$[\varepsilon_X, \varepsilon_Y]$(pixel)	Corr. Acc.(%)
Cross	86.21	11.63	2.16	[1.931, 1.927]	—
Rhombus	85.83	12.57	1.60	[2.056, 2.051]	—
Sandglass	87.49	11.64	0.87	[1.816, 1.821]	—
Whole Pattern	86.33	11.06	2.61	[2.013, 2.043]	91.23

TABLE I
THE QUANTITATIVE EXPERIMENT RESULTS

surface with bare-hand (Fig. 8 (c)). From an image alone, say of a finger on top of a table surface, one cannot tell whether the finger is actually touching the table surface or not. The case of a finger hanging in air, and the case of a finger touching the table surface, could both produce the same image to the camera. By incorporating the structured light invisible embedded into the projection, 3D acquisition can be made possible, and contact identification and finger movement recognition should be more readily tackled[2]. It is possible to convert any textureless light color plane (table-surfaces, whiteboards or walls) to be a touching sensitive screen, providing more natural and flexible interface for bare-hand human-robot interaction.

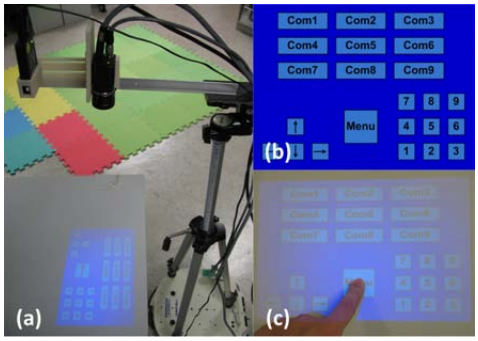


Fig. 8. Some 3D sensing results

VI. CONCLUSION

We have described a novel system of embedding imperceptible structured codes into regular video projection that strikes the balance between imperceptibility and detectability of the codes. Through precise projector-camera synchronization, structured codes consisting of three primitive shapes are embedded into the regular projection, in a way that is imperceptible to the user but extractable by a camera (through the "difference image" between successive images). The disturbances caused by external noise make it difficult to retrieve the codes by the region segmentation approaches adopted in general structured light systems. Instead of segmenting the codes, specially trained classifiers are employed to detect and identify them. To raise the robustness of code extraction, large Hamming distance is adopted in spatial coding. Even if some bits are missed or wrongly decoded, the correct correspondence between the projection panel and the image plane could still be arrived at correctly for structured light sensing. Extensive experimentation shows that the method is a feasible and promising. Some examples are given to demonstrate the potential applications in robotic system, such as sensing surrounding environment and natural human-robot interaction.

[2]We have implemented fingertip touching detection method under invisible codes embedded illumination, but this is detailed in other paper due to space constraints.

VII. ACKNOWLEDGMENT

This work is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing & Interface Technologies.

REFERENCES

[1] H. Park, B. Seo and J. Park, Subjective evaluation on visual perceptibility of embedding complementary patterns for nonintrusive projection-based augmented reality, *IEEE Transactions on Circuits and Systems for Video Technology*, 20(5):687-696, 2010.
[2] O. Bimber, D. Iwai, G. Wetzstein and A. Grundhöfer, The visual computing of projector-camera systems, *ACM SIGGRAPH 2008 classes*, pp. 1-25, 2008.
[3] Z. Song and R. Chung, Use of LCD panel for calibrating structured-light-based range sensing system, *IEEE Transactions on Instrumentation and Measurement*, 57(11):2623-2630, 2008.
[4] D. Fofi, T. Sliwa and Y. Voisin, A comparative survey on invisible structured light, *Proceedings of Machine Vision Applications in Industrial Inspection XII*, pp. 90-98, 2004.
[5] J. Salvi, S. Fernandez, T. Pribanic and X. Llado, A state of the art in structured light patterns for surface profilometry, *Pattern Recognition*, 43(8):2666-2680, 2010.
[6] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs, The office of the future: A unified approach to image-based modeling and spatially immersive displays *Proceedings of SIGGRAPH 98*, pp. 179-188, 1998.
[7] D. Cotting, M. Naef, M. Cross and H. Fuchs, Embedding imperceptible patterns into projected images for simultaneous acquisition and display, *Proceedings of The IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 100-109, 2004.
[8] T. Etzion, Constructions for perfect maps and pseudorandom arrays, *IEEE Transactions on Information Theory*, 34(5):1308-1316, 1988.
[9] C. Albitar, P. Graebling and C. Doignon, Robust structured light coding for 3D reconstruction, *Proceedings of IEEE 11th International Conference on Computer Vision*, pp. 1-6, 2007.
[10] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* vol.1, pp. 511-518, 2001.
[11] R. Lienhart, A. Kuranov and V.Pisarevsky, Empirical analysis of detection cascades of boosted classifiers for rapid object detection, *In DAGM 25th Pattern Recognition Symposium*, pp. 297-304, 2003.
[12] H. Park, M. Lee, B. Seo, Y. Jin and J. Park, Content adaptive embedding of complementary patterns for nonintrusive direct-projected augmented reality, *HCI international*, pp. 132-141, 2007.
[13] A. Grundhofer, M. Seeger, F. Hantsch and O. Bimber, Dynamic adaptation of projected imperceptible codes, *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 1-10, 2007.
[14] ARRICK Robotics, http://www.arrickrobotics.com/.
[15] TI Pico Development Kit, http://www.ti.com/tool/dlp1picokit/.