2022 IEEE International Conference on Robotics and Automation (ICRA)
May 23-27, 2022. Philadelphia, PA, USA

SpecTac: A Visual-Tactile Dual-Modality Sensor Using UV Illumination

Qi Wang*, Yipai Du* , and Michael Yu Wang, *Fellow, IEEE*

2022 IEEE International Conference on Robotics and Automation (ICRA) | 978-1-7281-9681-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRA46639.2022.9812348

Abstract— **Perceiving the dynamical environment both visually and tactilely is crucial for the survival of animals, and therefore, is considered of importance in robotics research. Recently, there has been an increasing interest in vision-based tactile sensors due to their high sensing resolution and robustness to environmental changes. However, almost all vision-based tactile sensors make only partial use of the camera, specifically, only when contact occurs, and stay idle at other times, which results in a waste of the camera information bandwidth. In this paper, we propose a new visual-tactile dual-modality sensor called SpecTac, which can visually inspect the environment and make tactile observations. The main novelty of the sensor is the use of ultraviolet (UV) LEDs and randomly distributed UV fluorescent markers. When the LEDs are on, those markers will be bright and can easily be distinguished and tracked from the background. Besides, by controlling the on and off of the UV LEDs, due to the switchable visibility of those markers, the sensor will switch between visual and tactile sensing mode. The qualities of tactile and visual perception are evaluated quantitatively by force estimation, visual triangulation and visual feature matching. By combining both modalities into one compact sensor, the information from the camera is better utilized, and it is hoped that the sensor will achieve more flexibility in the motion of the robot arm, especially in tasks where the workspace is narrow.**

I. INTRODUCTION

For autonomous robots, it is critical to perceive dynamic environments and make appropriate responses. By vision-guided motion and tactile-based control, robots can track moving objects [1] and avoid grasp failure [2]. Furthermore, the fusion of vision and touch can enable robots to complete more complex and dexterous manipulations.

Visual perception based on cameras plays a crucial role in robotic manipulation tasks. It is difficult for robots without cameras to localize objects and track them in closed-loop manipulation. However, self-occlusion can sometimes be a problem with a single camera. Multiple cameras can also work in a cooperation scheme for manipulation: the fixed camera localizes objects in the environment and observes the scene, while the movable camera tracks the objects relative to the end effector [3]. Increasing the visual perception sensors can help to do more dexterous manipulation tasks.

Recently, vision-based tactile sensors using digital imaging through a transparent gel have been developed, such as GelSight [4], GelSlim [5], and Digit [6]. These sensors have achieved outstanding sensing resolution and robustness to the environment. They can reconstruct force information with a camera from marker motion and photometric stereo, which can increase the flexibility and robustness of dexterous manipulations [7], [8]. However, due to the imaging used, the cameras in the sensors are usually sealed for illumination constancy. Hence, they only matter when contact occurs but keep capturing the same unchanged image for the rest of the time. The information bandwidth is therefore largely wasted. Yamaguchi *et al.* [9] proposed FingerVision sensor, which works without an opaquely coated surface. However, the visual perception ability is limited to proximity sensing. Abad *et al.* [10] introduced UV markers, but the use of those markers is only to avoid leaving holes in the RGB image.

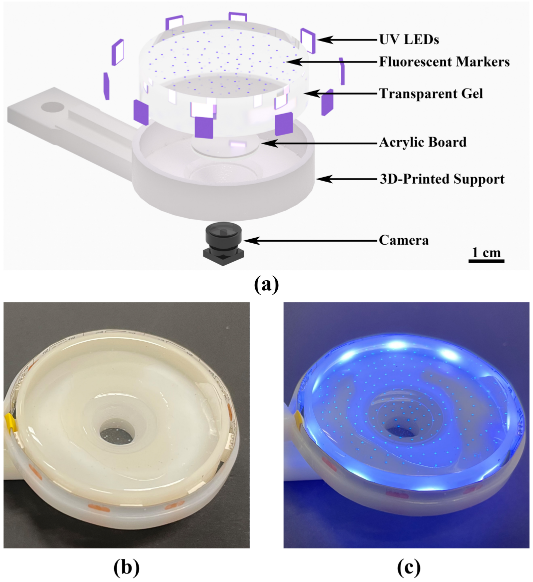


Fig. 1. (a) Schematics of the design for the visual-tactile dual-modality sensor. The appearance (b) without LEDs and (c) with LEDs lit up.

To tackle this problem, we present a visual-tactile dual-modality sensor, called SpecTac, as shown in Fig. 1. There are six main parts of the sensor: a camera, 3D-printed support, acrylic board, cylindrical transparent gel, randomly painted fluorescent markers on the gel, and LEDs that can emit UV light. It has both visual and tactile modalities. When the LEDs are off, the camera can view the scene through the transparent gel. When the LEDs are on, the fluorescent markers will illuminate, and thus, the camera can capture them as bright blue dots on the image. These dots are easy to distinguish, detect and track from the surroundings. Therefore, by controlling the LEDs turning off and on, the sensor can switch its modality in visual and tactile sensing. By using a single camera through time-division multiplexing (TDM), the SpecTac sensor integrates the functionalities of both wrist-mounted camera and vision-

Qi Wang, Yipai Du (corresponding author) and Michael Yu Wang are with the Hong Kong University of Science and Technology (e-mail: {qwangcl, yduaz, mywang}@ust.hk). Michael Yu Wang is also with HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen.

* Equal contribution.

based tactile sensor. A comparison of the SpecTac sensor and other sensors for robotic manipulation is listed in Table. I. It can do manipulation tasks like inspecting the scene, localizing the target objects, and grasping dexterously using tactile information.

This paper is structured as follows: In Sec. II, related works about tactile sensing and visual sensing, especially in the field of robotic manipulation are introduced. In Sec. III, the design of the hardware and software architecture are described in detail. In Sec. IV, normal and shear force estimation based on the displacement of markers are used to evaluate the tactile perception quality of the sensor. The visual perception is evaluated by 3D triangulation, SIFT [11] feature detection and matching. In Sec. V, a summary of the work and future research directions are discussed.

TABLE I
THE COMPARISON OF DIFFERENT SENSORS FOR THE ABILITY OF DOING TACTILE AND VISUAL TASK

Sensors	GelSight[4], GelSlim [5], Digit[6]	FingerVision[9]	Camera	SpecTac
Tactile tasks	✓	✓	✗	✓
Visual tasks	✗	only proximity sensing	✓	✓

II. Related Works

A. Tactile Sensing and Vision-based Tactile Sensors

Tactile sensing is an essential ability for animals to perceive and react to their environment. Therefore, it has been studied for years in robotics research [12]. Resistive, capacitive, and piezoelectrical transduction interfaces are used in traditional tactile sensors. However, when the area to be measured is large, such sensors generally suffer from sensitivity to external changes (e.g., temperature fluctuations and electrical interference) and complicated wiring due to electric single point signals involved [13]. In recent years vision-based tactile sensing approaches are prospering, owing to their better sensing resolution, easy manufacturing method, robustness in harsh settings, multi-axial measuring capability, and simple multiplexing peripherals. Furthermore, current developments in digital cameras not only make their application low-cost and simple-to-use, but also synergize tactile sensing with computer vision and deep learning, allowing the transfer of visual perception knowledge to tactile perception. Because of the higher resolution and distinct representation of vision-based tactile sensors, processing algorithms for classical tactile sensors cannot be directly transferred to them.

Vision-based tactile sensing often uses a soft surface that is responsive to contact. When it deforms, the change of shape is captured by the camera in the 2D image. This idea was first implemented by Kamiyama *et al.* [14] with a color CCD camera recording the positional center of mass variation in the markers. In the last two decades, other approaches to solving tactile sensing have emerged. The GelSight sensor uses photometric stereo to reconstruct the contact depth map from three color illuminations [15]. Dot markers were later added to the GelSight sensor to allow for measurements not only in the normal direction but also in the shear direction [4]. While most other methods used opaque sensing surfaces, Yamaguchi *et al.* [9] chose not to cover the imaging system completely for isolated illumination environment, but left the silicone gel transparent to provide proximity vision ability. GelSlim sensors use reflective mirrors [5] and a shaping lens [16] to reduce the thickness of the tactile sensor so that they are compact enough to be used as grippers to squeeze between objects. OmniTact [17] uses five cameras inside a single sensor to provide a wide field of view and multi-directional sensing ability, which is better suited for fingertip touch sensing. *Soft-bubble* grippers use an air-inflated structure to make the contact surface more compliant to the grasped object with a large friction force [18]. A compact ToF (Time of Flight) depth sensor is used in *soft-bubble* to enable contact shape reconstruction and object pose estimation. The TacTip sensor [19] uses an array of pins on the inside of the sensing surface to form a compliant flesh-like structure, with the aim to mimic the principle of transduction of tactile stimuli for human skin. Sferrazza *et al.* [20] adopted randomly scattered particles as tracking targets and combined dense optical flow and a neural network for accurate force estimation. DIGIT [6] miniaturizes the form factor of vision-based tactile sensors so as to be mountable on a multi-fingered Allegro hand for dexterous in-hand manipulation.

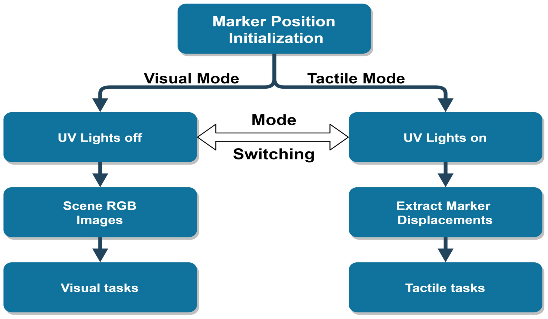


Fig. 2. Working principle for SpecTac: The left branch is for visual mode and the right branch is for tactile mode. The sensor can switch the modes by the on and off of UV lights.

B. Visual and Tactile Sensing for Robotic Manipulation

Visual perception is the most common modality in robotic manipulation tasks. Researchers have utilized depth cameras like Microsoft Kinect [21] and Intel RealSense [22], or RGB cameras which can be either installed as stationary or on the wrist of the robot hand [23] for a close and detailed view. Visual perception alone can provide 3D semantic information and an object's position, orientation, depth, the most basic prior knowledge a robot needs to interact with its environment.

On the other hand, tactile sensing provides subtle but crucial information that originates from the contact surface which is not possible for external vision systems to capture. Various extraction and interpretation techniques have been developed based on extracted tactile signals, such as force estimation [2], [20], geometry estimation [24], slip detection

[25], predicting slip before it happens [2], grasp stability evaluation and regrasp planning [26], [27], contact event classification [28], and extrinsic contact sensing [29].

The tactile sensing and visual sensing modalities can complement each other and can be connected [30]. Both modalities combined improve grasping performance [26], [31], [32], but more importantly, the fused information helps robots to learn faster with fewer data required [33]. The key observation is the different times when the two modalities become useful. During a robotic manipulation task, the vision gives a distant view, while the tactile sensing provides local detail when a contact of the robot and the object happens [34]. This inspires us to combine the two modalities using a single camera at different times. It is hoped that this compact design could help improve the flexibility of dexterous manipulation and ease the process of hand-eye calibration.

III. System and Methods

In this section, we give a detailed description of the hardware and software architecture of our proposed sensor.

```
Algorithm 1: Tactile Feature Extraction
Input: Initial tactile image $\mathbf{t}_0$,
       Tactile image sequence $\mathcal{T} = \{\mathbf{t}_i | i = 1, 2, ...\}$
Output: Tactile feature displacement grid $\mathcal{F} = \{\mathbf{f}_i | i = 1, 2, ...\}$
$\mathbf{Mask}_0$ =Thresholding(DoG(RGB2YUV($\mathbf{t}_0$)));
$\mathbf{Mask}_0$ =Morphology($\mathbf{Mask}_0$);
$\mathbf{P}_0$ =Blob detection in $\mathbf{t}_0$ at $\mathbf{Mask}_0$;
for $\mathbf{t}_i$ in $\mathcal{T}$ do
    $\mathbf{Mask}_i$ =Thresholding(DoG(RGB2YUV($\mathbf{t}_i$)));
    $\mathbf{Mask}_i$ =Morphology($\mathbf{Mask}_i$);
    $\hat{\mathbf{P}}_i$ =Blob detection in $\mathbf{t}_i$ at $\mathbf{Mask}_i$;
    $\mathbf{f}_i = [\ ]$;
    for $p^k_{i-1}$ in $\mathbf{P}_{i-1}$ do
        $n_1, n_2$ =kNN($p^k_{i-1}$, $\hat{\mathbf{P}}_i$) (k=2);
        if distance($n_1$, $p^k_{i-1}$)$<0.5*$distance($n_2$, $p^k_{i-1}$) and
          distance($n_1$, $p^k_{i-1}$)$<8$ then
            $p^k_i = n_1$;
            $\mathbf{f}_i$.append($p^k_i - p^k_0$);
        else
            $p^k_i = p^k_{i-1}$;
        end
    end
    $\mathbf{f}_i$ =griddata($\mathbf{f}_i$)
end
```

A. Sensor Design and Fabrication

To fabricate our SpecTac sensor, we put a piece of Solaris™ silicone gel (about 10 mm thickness), which is soft and transparent on top of a USB RGB camera (HBV-1466M12 S1.0) running at 30 FPS equipped with a 160° wide-angle lens. Around 130 randomly distributed fluorescent markers are painted on the surface of the transparent gel by hand using a small brush as tactile tracking features, resulting in a marker density of $8.2/\text{cm}^2$. The markers can be made smaller and the density can also be increased with advanced manufacturing techniques, which is out of the scope for proving the concept in this paper. The markers are barely visible under natural light conditions but can emit blue light under UV illumination. Another Solaris™ layer about 0.5 mm thick is put on the top of the sensor surface for protection during contact. The sensor body is 3D printed using SLA material as the support structure for the parts. Surrounding the gel are 10 LEDs (5 V/DC) that can emit UV light when powered on.

The sensor is designed to have two working modes: visual and tactile. Under visual mode, the LEDs are switched off and the camera can see the outside world clearly through the transparent gel. On the other hand, under the tactile mode, the UV LEDs are powered on and the markers will be lit up so they can be tracked using the SimpleBlobDetector in OpenCV [35]. Moreover, some light of LEDs can be reflected by the silicone gel-air interface, and the LEDs' lights are visible due to the broad spectrum. These internal reflected lights are stable and constant, which can help to neutralize dynamic environmental light changes and stabilize the blob detection process. With the tracked marker motion, one can do many different types of tactile processing including force estimation, contact event classification, grasping evaluation and regrasping etc., as discussed in Sec. II. An Arduino Uno is used to turn on and off the LEDs to switch between the visual and tactile mode. The working principle is summarized in Fig. 2.

B. Tactile Feature Extraction

Under tactile mode, the UV LEDs are powered on so the markers are visible to the camera for contact deformation tracking. The tactile processing consists of three parts: image pre-processing, blob detection and marker displacement extraction.

Image Pre-processing In the beginning, the camera exposure value is set to -6 for constant image brightness. The RGB images are firstly converted to the YUV color space. Because the markers are semi-transparent, their apparent colors are mixed with the external scene. To enhance the tactile marker features with a specific size and color (in YUV color space), a difference of Gaussian (DoG) filter is applied with standard deviations of 50 and 60, respectively. Then the locations where the YUV pixels lie within the upper bound $(23, 18, 0)$ and lower bound $(-5, 3, -14)$ are selected as the regions to look for blob markers. These masked regions undergo morphological closing and opening operations to close the small gaps and remove salt and pepper noise.

Blob Detection At the masked regions, the SimpleBlobDetector in OpenCV is used to detect the location of the fluorescent markers. The blob detection parameters are optimized manually according to the characteristics of the painted markers for the sensor to obtain more accurate results. The goal for pre-processing and blob detection is to maximize the number of tactile markers detected. Some markers may not be detected due to external light disturbance, while features in the external environment may be mistakenly detected in this process. Hence a way to update the displacements for the missing markers as well as to filter out the mistaken features that are not from the tactile markers is needed in the next step.

Marker Displacement Extraction In the initialization stage, all the tactile markers need to be accurately detected for tracking. For every new incoming frame, its detected blob points are compared with the locations of the blob points from the previous frame. The correspondences are established by finding the nearest neighbors of every previously tracked point from the newly detected points. Moreover, a correspondence is considered valid only if the Euclidean distance between the blob points is less than 8 pixels, and the distance from the points to their nearest neighbors in the new frame is less than half of the distance to their second-nearest neighbor. If one previously tracked point fails to find a proper match in the current image, it is temporarily disabled and its location remains unchanged. Finally, the displacements are calculated for all the enabled points in the current frame compared to their initial position. For consistency of different marker displacements at all time, the 2D displacements of the sparse points are interpolated using *griddata* from scipy package [36] into a 12×12 regular grid.

The complete algorithm for tactile feature extraction is summarized in Algorithm. 1. It is able to run at 30 FPS on a personal laptop hence is fast enough to process one frame before the next frame from the camera arrives.

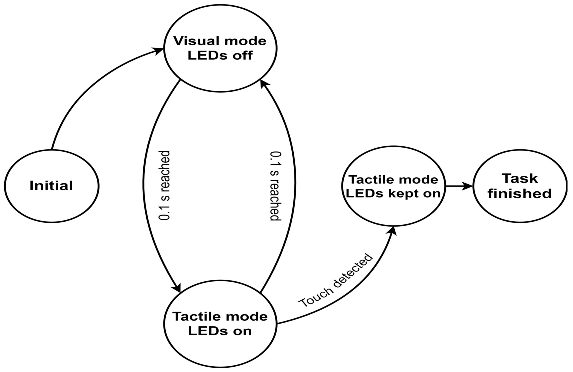


Fig. 3. The state transition graph for the visual-tactile modality switching in a typical manipulation task.

C. Modality Switching Strategy

For the visual-tactile system to function, the UV LED lights flicker under the control of Arduino Uno so that the sensor constantly switches between the two perception modalities. The first step is to initialize the positions of the markers as a reference following the process in Sec. III-B. Before reaching the desired object, the main function of the sensor is to observe and inspect the scene, and the UV LEDs are powered on at a duty cycle of 0.1 seconds in every 0.2 seconds, whereas the visual perception is done in the remaining 0.1 seconds. The choice of this particular duty cycle is reasoned in Sec. IV-D. In the 0.1 seconds when the LEDs are powered on, the positions of the detected markers are compared with their initial positions. For robustness consideration, if the average displacement of the markers is greater than 2 pixels (which is easy to be achieved for a real contact but not possible to result from blob tracking error in sub-pixel accuracy), the sensor is considered to have made contact, so the sensor enters tactile mode and focuses on the tactile signal acquisition until the manipulation task is finished. The modality switching strategy is shown as a state graph in Fig. 3. A practical example can be found in the supplementary video.

In this section, experimental details are given and results are presented to show that the sensor can work as a tactile sensor while in contact with an object, and as a regular RGB camera for environment observation when no contact is made. With the tracked marker motions, various types of tactile sensing have been demonstrated to be efficacious by prior research, e.g., entropy-based slip detection [2] and learning-based contact event classification [28]. Hence, for simplicity, we ease the experimental burden by choosing normal and shear force estimation to evaluate the tactile processing performance since it is the fundamental type of tactile sensing. All the aforementioned previous works on tactile processing are directly transferable to the SpecTac tracked markers. For visual sensing, 3D triangulation and SIFT matching are conducted to evaluate the visual sensing quality. More visual perception tasks with the sensor are possible, such as 3D reconstruction, scene segmentation etc., and these are left as future work.

IV. Experiments and Results

A. Normal and Shear Force Estimation

TABLE II

Force Estimation Evaluation

Model	Error (N)	F_x (Shear)	F_y (Shear)	F_z (Normal)
Linear Regression	Mean	0.10	0.078	0.34
	Std	0.092	0.074	0.29
SVR (RBF)	Mean	0.11	0.11	0.25
	Std	0.16	0.15	0.32
MLP	Mean	**0.056**	**0.052**	**0.12**
	Std	**0.052**	**0.051**	**0.10**

The force estimation is based on a data-driven approach. We 3D print **6** spherical indentors with radii 8 mm, 10 mm, 12 mm, 15 mm, 20 mm, 25 mm respectively to make contact with the SpecTac sensor. The indentors are mounted on an ATI Mini 27 Titanium F/T sensor to obtain the 3-axis force ground truth. The SpecTac sensor is mounted on a 3-axis linear stage to automate the data collection process, as shown in Fig. 4(a). Each indentor makes contact at $\mathbf{3} \times \mathbf{3}$ grid cells distributed on the gel surface of the SpecTac sensor. At each location, there are **5** indentation levels with 0.5 mm step size. At each indentation level, 2 extra steps to move up, down, left, right 0.3 and 0.6 mm are made to introduce shear force. This results in **9** data points at each location per indentation level. Therefore, there are in total $\mathbf{6} \times \mathbf{3} \times \mathbf{3} \times \mathbf{5} \times \mathbf{9} = 2430$ data points being recorded. The shear force ranges from -3 N to 3 N while the normal force ranges from 0 N to 15 N. 70% of the data are randomly selected for the training purpose and the remaining 30% for testing. As for the regression models, simple linear regression, SVR (support vector regression) with RBF kernel and MLP (multilayer perceptron) are adopted. The MLP is designed to have 4 hidden layers with 200 hidden units each. Table. II shows the comparison of different models on the precision of force estimation. It can be seen that the MLP model outperforms

TABLE III

THE AVERAGE NUMBER OF GOOD SIFT MATCHES IN EACH SCENE

Images	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Scene 6	Scene 7	Scene 8
SpecTac sensor with mask	0.66	0.66	0.74	0.75	0.77	0.72	0.69	0.70
Camera with mask	0.94	0.93	0.92	0.98	0.89	0.92	0.94	0.94
Camera without mask	1.0 (95.8)	1.0 (165.6)	1.0 (81.3)	1.0 (21.3)	1.0 (124.3)	1.0 (90.4)	1.0 (106.6)	1.0 (122.5)

the other models by a large margin. It is generally more challenging to estimate the normal force than the shear force, mainly due to the depth ambiguity in monocular settings. It is worth noting that the inherent noise of the F/T sensor we used to collect ground truth is 0.03 N for shear force and 0.06 N for normal force. Hence the force precision of the MLP model is rather satisfactory considering the inherent limitation in the ground truth data.

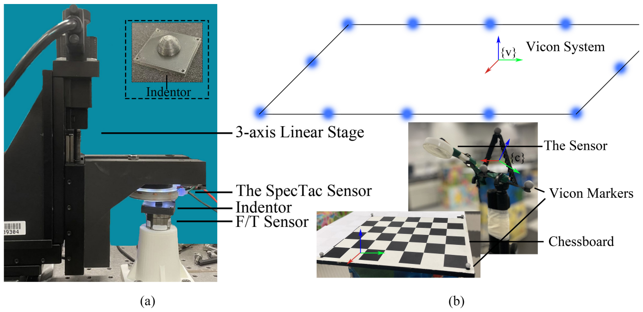


Fig. 4. (a) The experiment setup of force dataset collection.(b) The setup of 3D triangulation. The sensor used is the SpecTac or a regular camera. The blue circles on the top represent the Vicon cameras.

B. Visual Calibration and 3D Triangulation

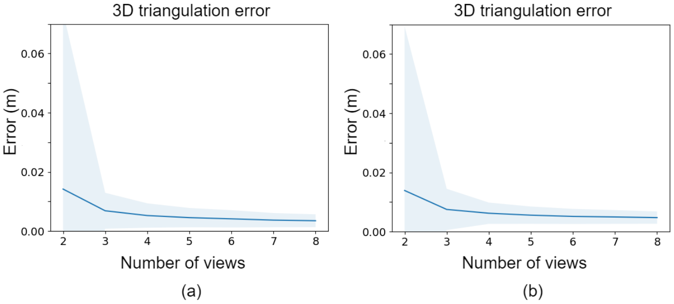


Fig. 5. The absolute triangulation error (distance between estimation and ground truth) in 3D space. The curves show the mean error of triangulation with 2 to 8 views using random samples taken from the test dataset. The shaded parts are above and below one standard deviation. (a) Triangulation with the camera directly. (b) Triangulation with the SpecTac sensor and the LEDs off.

This section illustrates the calibration process of the RGB camera and presents a multi-view triangulation experiment to evaluate the effectiveness of the intrinsic/extrinsic calibration of the camera. A chessboard with 6×5 corner points (shown in Fig. 4(b)) is used for the experiment. The spacing of the neighboring corner points is 0.08 m. With the aim to compare the image quality regarding camera calibration and 2D to 3D triangulation, the same process is carried out using the SpecTac sensor (LEDs off) and the original camera. Both the board and cameras are equipped with Vicon markers so the 3D position of the corner points and 6D pose of the Vicon markers mounted on the camera can be accurately tracked in the Vicon coordinate frame $\{v\}$. The camera's external coordinate frame $\{c\}$ is represented by the Vicon markers mounted on it. The aim of the camera calibration is to find the geometric transformation from $\{c\}$ to the image frame $\{i\}$. With the Vicon tracking system, the transformation $\mathbf{T}_{cv}$ from $\{v\}$ to $\{c\}$ is obtained directly. Each chessboard corner point $\mathbf{P}_v$ in the Vicon frame can be transformed into the camera frame by

$$\mathbf{P}_c = \mathbf{T}_{cv} \cdot \mathbf{P}_v. \tag{1}$$

Then, the aim of the camera calibration process is to find the proper intrinsic $\mathbf{K}$, extrinsic $\mathbf{R}|\mathbf{T}$ and distortion parameters such that

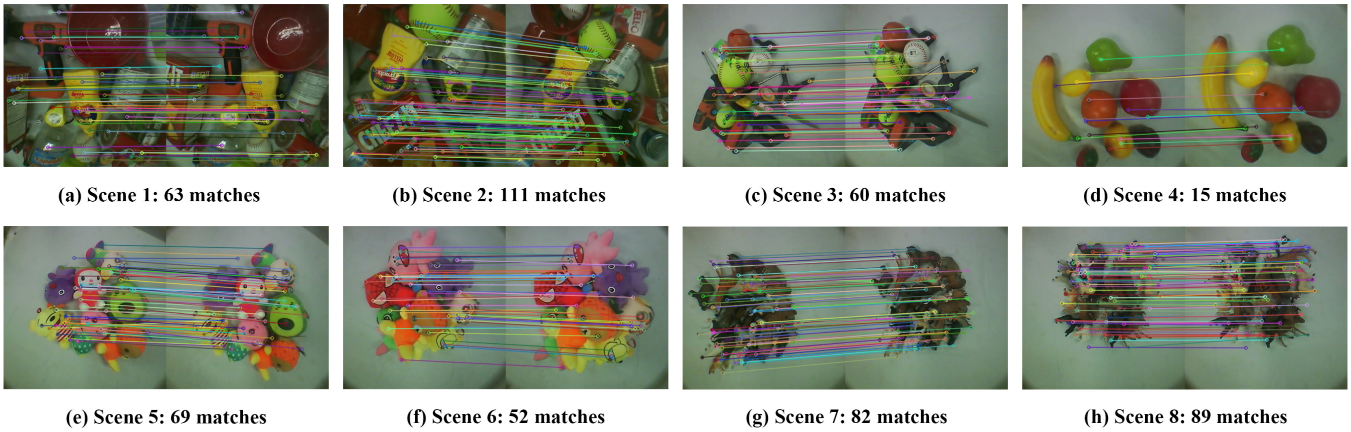
$$\mathbf{p} = \mathbf{K} \cdot \boldsymbol{f}(\mathbf{R} \cdot \mathbf{P}_c + \mathbf{T}), \tag{2}$$

where $\boldsymbol{f}(\cdot)$ is the undistort function and $\mathbf{p}$ is the pixel coordinate in $\{i\}$. In practice, we fix the chessboard and move the camera so different locations of $\mathbf{P}_c$ are obtained. We take 70 images with the original camera and 70 with the SpecTac sensor (LEDs off). The images are cropped to 320×320 to the region of interest. For each group of images, 30 images are used for camera calibration and 40 images for testing purposes. The calibration utilizes the calibrateCamera and solvePnPRansac functions from OpenCV [35].

To test the calibration accuracy, a multi-view triangulation experiment is conducted. The triangulation is that given $\mathbf{K}$, $\mathbf{R}$, $\mathbf{T}$, $\boldsymbol{f}(\cdot)$ and multiple $\mathbf{T}_{cv}$ and $\mathbf{p}$ from different views, estimating the location of $\mathbf{P}_v$. For simplicity, a linear method of triangulation is adopted by solving an SVD problem [37]. The experiment is done by drawing 2 to 8 random samples from the test dataset, and for each number of views, 1000 triangulations are done with randomly sampled image input to obtain the absolute triangulation error statistics. As shown in Fig. 5, the average triangulation error stabilizes to less than 0.01 m from 3-view triangulation and above for both the original camera and the SpecTac sensor. The SpecTac sensor exhibits a larger triangulation error, but the defect is not significant. The considerable error for the 2-view triangulation may be due to the fact that there exist some image pairs that are taken too close to each other, resulting in large uncertainty in the triangulation. Considering the average distance to the chessboard is over 1.0 m and the cheap camera module used, the 3D triangulation error is relatively small (less than 1% at 1.0 m). Moreover, adding the tactile layer to the camera does not introduce a significant downgrade of the triangulation performance.

C. Visual Matching with SIFT Descriptor

In order to evaluate the visual quality of the SpecTac sensor, a SIFT feature detection (with OpenCV default parameters) and matching experiment is done to quantify



(a) Scene 1: 63 matches

(b) Scene 2: 111 matches

(c) Scene 3: 60 matches

(d) Scene 4: 15 matches

(e) Scene 5: 69 matches

(f) Scene 6: 52 matches

(g) Scene 7: 82 matches

(h) Scene 8: 89 matches

Fig. 6. Example SIFT matching results with the SpecTac sensor in each scene. For most cases the number of good matches are over 50. In scene 4 the amount of match is small due to being too simple and textureless.

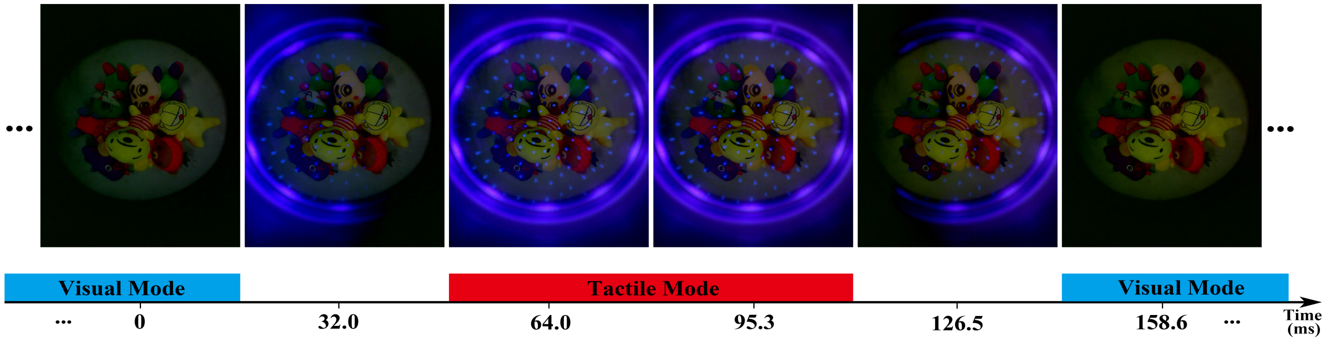


Fig. 7. Consecutive image sequence when the SpecTac sensor changes modes. The horizontal axis represents the relative time stamp of every frame. The red block indicates that the images within this region are used as tactile image; likewise, the images within the blue regions are used as visual image. The second and fifth images are half illuminated and are abandoned in practice.

the number of good matches in image pairs captured by the sensor in visual mode. Although the markers are semi-transparent, the light passing through still undergoes some distortion. Hence, the locations of the markers are masked out in the SIFT feature detection. A separate camera in the same pose as the sensor captures images for comparison. We set up 8 scenes with daily objects and take 3 image pairs in each scene using both the sensor (LEDs off) and the camera. The average numbers of good matches in each scene are listed in Table. III. For clearness of comparison, the average number of matches is normalized by the number of good matches with the unmasked camera in each scene, while its absolute value is in the parentheses. Note that the masked area takes 3.3% of the total image area, but in general, masking those areas introduces a slightly larger match decrease. Nevertheless, it is still recommended to mask out those regions for the SpecTac sensor for better accuracy in the matching process in case the distortion caused by tactile features is detected as blobs. The number of remaining matches for the SpecTac sensor after masking is still applicable. Some example SIFT matching results of the SpecTac sensor are shown in Fig. 6.

D. Switching Time of the Sensor

When the mode switches, due to the rolling shutter camera and the time delay to control the LEDs, the captured images will not be responsive instantly. Fig. 7 shows the image sequence when the mode switches. It can be seen that only one image on the boundary between the two modes is half bright and half dark, which makes the image inappropriate for either of the two modes. Therefore, when the mode switches, there is one image frame being dropped intentionally in order to avoid this ambiguity. This will not influence much in the 30 FPS camera system.

V. Conclusion

In this work, a new visual-tactile dual-modality sensor called SpecTac is proposed. With the use of UV fluorescent markers, visual and tactile perception can be done in one compact sensor. The quantitative experiments on visual perception show that adding the tactile layer to the original camera does not impact the visual performance severely. The camera still suffices to accomplish visual tasks. The detection and tracking of markers under tactile sensing mode can be successfully used for normal and shear force estimation. With the tracked marker motion, many existing tactile processing techniques are possible, which shows great potential for visual-tactile fusion. We also propose a strategy for dynamic modality switching specifically for the SpecTac sensor to make maximal use of the camera information bandwidth. Future research direction includes making the gripper design more suitable for this kind of visual-tactile sensing and visual-tactile cooperation in more sophisticated scenarios.

VI. Acknowledgment

This work was supported by the Hong Kong Innovationand Technology Fund (ITF) under Grant ITS/104/19FP, and supported in part by the Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYB-2020083). We thank Zicheng Kan for the help with 3D printing and Tania Leigh Wilmshurst for proofreading.

References

[1] J. Huang, F. Zhang, X. Dong, R. Yang, J. Xie, and W. Shang, "Vision-guided dynamic object grasping of robotic manipulators," in *2020 IEEE International Conference on Mechatronics and Automation (ICMA)*. IEEE, 2020, pp. 460–465.

[2] W. Yuan, R. Li, M. A. Srinivasan, and E. H. Adelson, "Measurement of shear and slip with a gelsight tactile sensor," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 304–311.

[3] G. Flandin, F. Chaumette, and E. Marchand, "Eye-in-hand/eye-to-hand cooperation for visual servoing," in *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, vol. 3. IEEE, 2000, pp. 2741–2746.

[4] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[5] E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson, and A. Rodriguez, "Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1927–1934.

[6] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer *et al.*, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[7] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *arXiv preprint arXiv:1910.02860*, 2019.

[8] C. Wang, S. Wang, B. Romero, F. Veiga, and E. Adelson, "Swingbot: Learning physical features from in-hand tactile exploration for dynamic swing-up manipulation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5633–5640.

[9] A. Yamaguchi and C. G. Atkeson, "Implementing tactile behaviors using fingervision," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. IEEE, 2017, pp. 241–248.

[10] A. C. Abad and A. Ranasinghe, "Low-cost gelsight with uv markings: Feature extraction of objects using alexnet and optical flow without 3d image reconstruction," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3680–3685.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] J. Tegin and J. Wikander, "Tactile sensing in intelligent robotic manipulation–a review," *Industrial Robot: An International Journal*, vol. 32, no. 1, pp. 64–70, 2005.

[13] H. Yousef, M. Boukallel, and K. Althoefer, "Tactile sensing for dexterous in-hand manipulation in robotics—a review," *Sensors and Actuators A: physical*, vol. 167, no. 2, pp. 171–187, 2011.

[14] K. Kamiyama, H. Kajimoto, M. Inami, N. Kawakami, and S. Tachi, "A vision-based tactile sensor," in *International Conference on Artificial Reality and Telexistence*, 2001, pp. 127–134.

[15] R. Li and E. H. Adelson, "Sensing and recognizing surface textures using a gelsight sensor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1241–1247.

[16] I. Taylor, S. Dong, and A. Rodriguez, "Gelslim3. 0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger," *arXiv preprint arXiv:2103.12269*, 2021.

[17] A. Padmanabha, F. Ebert, S. Tian, R. Calandra, C. Finn, and S. Levine, "Omnitact: A multi-directional high-resolution touch sensor," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 618–624.

[18] N. Kuppuswamy, A. Alspach, A. Uttamchandani, S. Creasey, T. Ikeda, and R. Tedrake, "Soft-bubble grippers for robust and perceptive manipulation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9917–9924.

[19] N. F. Lepora, "Soft biomimetic optical tactile sensing with the tactip: A review," *arXiv preprint arXiv:2105.14455*, 2021.

[20] C. Sferrazza and R. D'Andrea, "Design, motivation and evaluation of a full-resolution optical tactile sensor," *Sensors*, vol. 19, no. 4, p. 928, 2019.

[21] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[22] L. Keselman, J. Iselin Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel realsense stereoscopic depth cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1–10.

[23] "Wrist camera." [Online]. Available: https://robotiq.com/products/wrist-camera

[24] S. Dong, W. Yuan, and E. H. Adelson, "Improved gelsight tactile sensor for measuring geometry and slip," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 137–144.

[25] K. Van Wyk and J. Falco, "Slip detection: Analysis and calibration of univariate tactile signals," *arXiv preprint arXiv:1806.10451*, 2018.

[26] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300–3307, 2018.

[27] F. R. Hogan, M. Bauza, O. Canal, E. Donlon, and A. Rodriguez, "Tactile regrasp: Grasp adjustments via simulated tactile transformations," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 2963–2970.

[28] Y. Zhang, W. Yuan, Z. Kan, and M. Y. Wang, "Towards learning to detect and predict contact events on vision-based tactile sensors," in *Conference on Robot Learning*. PMLR, 2020, pp. 1395–1404.

[29] D. Ma, S. Dong, and A. Rodriguez, "Extrinsic contact sensing with relative-motion tracking from distributed tactile measurements," *arXiv preprint arXiv:2103.08108*, 2021.

[30] W. Yuan, S. Wang, S. Dong, and E. Adelson, "Connecting look and feel: Associating the visual and tactile properties of physical materials," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5580–5588.

[31] J. Li, S. Dong, and E. Adelson, "Slip detection with combined tactile and visual information," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7772–7777.

[32] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?" in *Conference on Robot Learning*. PMLR, 2017, pp. 314–323.

[33] S. Wang, M. Lambeta, P.-W. Chou, and R. Calandra, "Tacto: A fast, flexible and open-source simulator for high-resolution vision-based tactile sensors," *arXiv preprint arXiv:2012.08456*, 2020.

[34] S. Wang, J. Wu, X. Sun, W. Yuan, W. T. Freeman, J. B. Tenenbaum, and E. H. Adelson, "3d shape perception from monocular vision, touch, and shape priors," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1606–1613.

[35] Itseez, "Open source computer vision library," https://github.com/itseez/opencv, 2015.

[36] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[37] J. Chen, D. Wu, P. Song, F. Deng, Y. He, and S. Pang, "Multi-view triangulation: Systematic comparison and an improved method," *Ieee Access*, vol. 8, pp. 21 017–21 027, 2020.