15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Hotspots Detection for Machine Operation in Egocentric Vision

Longfei Chen[1] Kazuaki Kondo[1] Yuichi Nakamura[1] Dima Damen[2] and Walterio W. Mayol-Cuevas[2]

1 Academic Center for Computing and Media Studies, Electrical and Electronic Engineering, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan

2 Department of Computer Science, University of Bristol Woodland Road, Bristol, BS8 1UB UK

E-mail: 1 yuichi@media.kyoto-u.ac.jp 2 Walterio.Mayol-Cuevas@bristol.ac.uk

Abstract

This paper introduces a novel idea of unsupervised hotspots detection from first person vision (FPV) records. The purpose is to gather typical patterns of machine operations based on touching or manipulating those hotspots and summarize the patterns as guides for operations such as online operating manuals. We chose sewing machine operation as an example and demonstrated that, a good performance of hotspots detection can be achieved by utilizing multiple features, especially touch and hand motion. More importantly, detected hotspots in both temporal and spatial locations matches well the positions of key components such as buttons, levers, and other important portions essential for operating the machine.

1. Introduction

With the development of consumer wearable devices, it is nowadays relatively straightforward to record vast and various types of data from everyday actions as lifelogs. Our research aims at automatically acquiring information in this context for the huge amount of daily experiences, for understanding the activities involved and ultimately to provide guidance help. Related to this purpose, inexpensive RGB-D camera has greatly enhanced the efficiency of visual sensing. Many state-of-the-art works have explored recognition of human daily activities captured by egocentric vision [7, 3, 2], most of which are focusing on daily activities such as opening the coffee jar or grasp a mug when preparing coffee or making a cake in kitchen scene [7]. These activities tend to interact with a variety of objects in common living scenes with a restricted manipulating complexity. One major approach to analyze these egocentric experiences is through hands-objects interactions (HOI). Rogez et al. [3] build an RGB-D egocentric dataset in describing fine-grained gasps, which is suggestive of hand pose, hand-object contacting points and contact force vectors greatly contribute to understanding HOI activities. Touch points are certainly strong clues to things been interacted and activities using them.

Another set of works have been concerned with detecting touch events by taking advantage of depth devices. Such as using stereo cameras [4] or a combination of fixed depth camera and a thermal camera [1]. Wilson [5] firstly utilizes a single fixed depth camera to sense touch on a tabletop. However, these background surface modeling approaches are quite difficult to meet our demands for egocentric environments due to rapid background change. Omni-Touch [6] extends touch sensing to wearable devices and built a system to create touch surface everywhere, while touch detection is limited to areas around fingertips and sensitive to approaching angle.

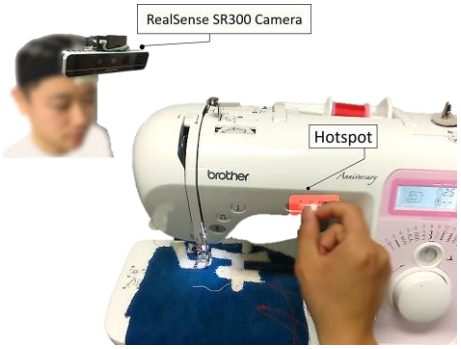

Figure 1.Machine operation experimental environment. Head-mounted RGB-D camera (top-left) records experiences while users are operating a sewing machine.

On the other hand, the wearer's point of attention is a strong clue of not only what is happening now, but importantly helpful to anticipate what action or object interaction will occur next. In [8], a regression model is built to estimate eye-gaze from head motion via an IMU in Google Glass. Damen et.al [2] discovered task-relevant objects by using eye fixations and linking gaze points to locations in the global scene. These attention-based methods appear useful to discover or predict important key events and objects in daily environments, as well as a way to reduce the amount of data that needs keeping.

In a similar spirit of discovering task-relevant objects, we hypothesize that important areas and sub areas on different kind of machines, e.g., a button, a lever, a switch, or a handle, can give us essential clues in understanding activities, i.e., how this part been manipulated and how the task has progressed. We define these crucial areas where everyone must interact with, which are comparatively "hotter" than other areas, as "hotspots".

In this paper, we present an unsupervised machine action hotspots detection method based on the combination of visual features. We chose a task of operating sewing machine, as shown in Figure 1, as a typical example. The task includes touching several small areas and a variety of hand-machine and hand-object interactions, e.g. touch, grasp, hold, push, pull, slide, rotate and cut etc.

The experimental result shows the potential of our approach for fine-grained hotspots detection in both temporal and spatial locations.

2. Framework

2.1 Machine Operation's Experiences Acquisition from Recorded FPV Videos

While operating a machine, people tend to remain in a stationary location relative to it. Both the head motion and gaze concentrate on it and are localized compared to during walking or a searching situation. There are a certain number of spatial areas that people often pay attention, touch, or manipulate. Most of them are important objects or interfaces such as buttons, levers, handles, which need to be touched for accomplishing a task. Hereafter, we consider such a spatial area as a "hotspot". The actions of touching or manipulating hotspots and their orders can thus be considered the essence of the task. But there are some difficulties and challenges: hotspots on machines can be sometimes small areas that are easily occluded by hands or human body; they may be close to each other; the actions to hotspots can be complicated.

This leads to a new application of computer vision; automatic hotspot detection and action classification or recognition. Detected hotspots and action contribute to summarize or analyze the task of machine operation.

2.2 Key Idea

The objectives of this research are, 1) automatic detection of hotspots, 2) extraction of interaction patterns relating to the hotspots.

Our idea is to use physical touches as primary clues to detect hotspots, while adding motion features (especially in depth) as subsidiary to enhance the performance. For this purpose, we take the advantage of egocentric vision which can provide less occlusion. Touches are detected based on the distance between hands and objects, and locations and frequency of touches are checked and integrated in the global map. Hand shapes in approaching or touching hotspots are used for discriminating hotspots and also categorizing the actions.

We chose a set of ordinary operations on a sewing machine as a testing case. Operations on sewing machines are composed of some steps that are not obvious and require certain prior knowledge and skill. A summary of the captured experiences would be useful as manuals for novice users, and their analysis useful for the repair engineers and designers. It can also be argued that capturing experiences with one sewing machine can be generalized to other sewing machines of similar size and interfaces. And intermediate operation steps could be useful to using even other types of machines such as a printer, a manipulation panel on a vehicle etc.

We also consider the integration of experiences taken from multiple users. This integration reduces dependence on personal differences and enables to look over the variations how people use a machine.

3 Touch Area Detection

For hotspots detection, local touch areas are detected from FPV frames and mapped to the global scene map.

3.1 Global Locations

Firstly, we need the global map on which detected touch areas are mapped and unified if possible. We expect that global map can be derived beforehand from the frames in which the whole machine surface appears and hands are not shown yet in recorded experiences. The frames for which a person is looking over the machine with the largest average distance from it are possible parts of the global map. They are stitched as follows. First, we may simply assume that each captured scene is a 2D plane, and the relationship between two planes can be described as a Homography transformation. We adopt SURF features and RANSAC to find corresponding paired points between two scenes in order to calculate the Homography transformation matrix. Although it causes some distortion on 3D shape, the distortion is not overly serious obstruction for locating the touch area and unifying them on the test case.

3.2 Hand and palm area

The extraction of exact palm area and its clear contour are important for obtaining accurate results of touch detection and shape description. Depth, color and size information are utilized to segment the hands areas. First, background is removed with depth information which refers to the distance. Then a chromatic histogram in HSV space of skin color for each user is built with several frames at the start of operation. The hand size is also obtained for the purpose of using as a constraint in detecting a hand, e.g., a hand size range in common hands working distance from a head mounted camera (20 cm to 100 cm).

The palm area is segmented out of the hand area based on the morphologic property. As shown in Figure 2, the arm area can be roughly regarded as a cylinder while the palm (or fist) is more circular. This method is also effective when there's no arm area appears in view.

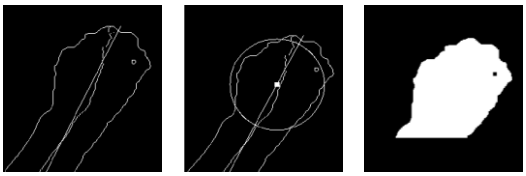

Figure 2.Palm detection relies on morphologic property. a) Find the midpoint of each row of a palm mask and fit with a line. b) From top to bottom of this line, create circles, calculate area ratio of hand in each circle. c) The circle with a maximum ratio is regarded as the palm area.

3.3 Find Touch in Depth

We propose a simple palm-oriented touch detection approach by comparing depth inside the palm and depth outside the palm, i.e., neighborhood depth along the palm contour. Figure 3 illustrate this method. Suppose that each small spatial window along the palm contour, mean depth inside hand is represent as S. If the outside neighborhood depth is smaller than $S + \Delta$, which means there are objects close to the palm, we regard that a touch has happened. There are random errors due to the camera's depth measurement accuracy (several millimeters in the range

of the SR300 camera). We thus set Δ as 7 mm in our experiment. Every detected touch area is indexed by its centroid (Cx, Cy) and radius (r). This method is relatively robust against hand shape and approaching angle changes in various manipulation circumstances, and it does not need a background model or a hand model that can be highly varying in shape. However, this simple approach may be affected by noises such as: i) casual touch by users unconsciously and that vanish out quickly, ii) misdetection caused by depth measurement errors, especially in the case that hands are close to surfaces.

Figure 3.Touch detection based on contour searching in depth (left). Detected touch areas are shown in red (right).

3.4 Hand Motion and Palm's Shape

3D hand motion is calculated by using position (x, y) in image plane and in depth z. In FPV, it is often observed that when hands are reaching towards something, the average depth value of hands increases; it will reach a local maximum at the moment contacting the surface. When touching finished, the average depth value of hands decreases. The depth value does not change in such "increasing-peak-decreasing" way in cases of hovering or lingering. We consider that the local maximum of hand depth values have a strong clue for touch, which may be used for filtering out misdetections. For palm shape features we use 2D contour points as proposed by [9]. Palms are rotated according to their principal axis of moment, and then the contour is sampled to 241 points.

4 Hotspot Detection and Interaction Pattern Classification

4.1 Hotspot Detection Algorithms

We can simply detect hotspots by finding areas with high frequency of touch in the global map. We call this simple method "frequency based approach (Fr)". For this purpose, the global map is divided into blocks with size $r \times r$, then, we calculate the accumulated frequency of touch in each block b_j ($j = 1, 2 \ldots m \times n/r^2$), where m and n are the size of global map. Blocks with frequency bigger than a threshold are detected as hotspots.

Some areas e.g. cloth plane, are touched many times with long duration, while some, e.g. needle button, are only touch once with a quick contact. However, the latter areas are also important in our task. For not only solving this problem but also detecting hotspots in temporal location, we apply a "temporal clustering approach (TC)" in a resolution of connected-area level as shown in Algorithm 1 below. The basic idea is as follows: although some important spots appear with low frequency, they appear locally high frequency in a small temporal window. We assume a small temporal window ω, if a certain touch area appears more that κ times in ω, it can be regarded as a valid touch. With this approach, we are able to maintain those spots in low global frequency yet are crucial ones.

Algorithm 1: Temporal clustering approach (TC)

Input: Reference frame R_0, window size ω, frequency threshold κ, hotspots number $n = 0$.
Output: Hotspots Θ_n and its start time St_n.

for $i = 1$ to end with step σ ($< \omega$) :
- a) Gather touch spots P_i in temporal window ω_i, create index frame M_i with P_i.
- b) For each connected-area C_j in M_i, clustering P_i to C_j, represent as P_{ij}. If number of $P_{ij} < \kappa$, erase corresponding area C_j.
- c) After checking all C_j, get new index frame M_i'.
- d) **if** area C_j' *in* M_i' is new appeared compare to R_{i-1}:
  n ++, $St_n = i$; $R_i = R_i + M_i$', $\Theta_n = C_j$'.
  else
  $R_i = R_i + M_i$'.

end

4.2 Interaction pattern

For each detected hotspot Θ_s in the spatial domain, the type of interaction is estimated. We first investigated the usage of the palm's shape. Although motions related to hotspots are also considered as a strong clue for indexing more complicated patterns, investigation is left for future work. To estimate interaction patterns, we first manually labeled more than a thousand palm shapes extracted from all recorded experiences beforehand and classified them into four catalogs: push, put (relax), rotate and slide. Then, a Random Forest classifier is trained with all shapes. In the actual hotspots detection and classification, palm shapes from all users are gathered for each hotspots, then they are classified by the above random forest classifier. The class with the largest number of classification is considered as the palm shape for manipulating the hotspots. This directly leads to the estimation of interaction type, e.g., push, slides, etc.

5 Experiment Result

Five experienced participants were asked to do a specific sewing machine operation while recording the FPV experiences. The task is determined by referring to the "Sewing Machine Operation Manual", and it is composed of 9 steps that require touches to 6 essential parts.

An Intel RealSense SR300 camera was used for capturing RGB-D images at 30 fps in resolution of 640×480 for both color and depth sources. For the frequency based approach (Fr) of hotspot detection, the block size is set as 20 pixels, and the frequency threshold is empirically set as 0.3 and 0.15 of their maximum for global palm and touch locations, respectively. For temporal clustering approach (TC), we find that depth measurement noises usually flash out quickly however the true touches usually last longer (at least 0.5s). Hence the within window frequency threshold κ is related to minimum valid touch time, it can be set around 0.3 of fps in our experiment. And the results are not sensitive to temporal window size because the windows have overlapping, we can set ω as

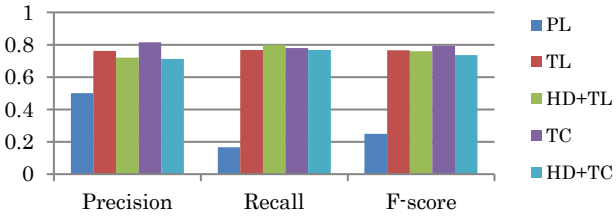

Figure 4.Average F-score of hotspots detection with 5 participants' experiences by different combination of features in ***Fr*** and ***TC*** approaches (Palm location and touch location is short as *PL* and *TL*, hand motion in depth is short as *HD*).

20 frames and the step σ to 0.25ω. For evaluating hotspot detection results, we use F-measure based on precision and recall. The ground truth is based on the task definition mentioned above, that is, where and how manipulates the sewing machine.

Figure 4 shows the average F-score of hotspots detection results for 5 participants. As our expectation, touch location (*TL*) is a strong feature in detecting global hotspots locations in Fr approach, while the palm location (*PL*) is not reliable for detecting locations. Because palms frequently appear and move continuously, we need stronger clues such as touch locations.

Another important feature we considered is the hand motion in depth (*HD*). We maintain all those touches nearby the peak locations in depth map without a frequency threshold. As illustrated in Figure 4, after adding *HD* to *TL*, the recall rate enhaced while the precision rate decreased. This shows *HD* is a good clue for crucial touches, however, casual touch noises are also included. On the other hand, *TC* approach has a better performance for precision and F-score. That is because *TC* approach can detect most of the crucial hotspots with either of high or low frequency ones, and it can also filter out some casual touches or misdetection. Additionally, *TC* based method provides a higher spatial resolution of hotspots as shown in Figure 5. However, the combination of *HD* and *TC* showed lower precision rate.

Table 1 shows the detailed result for the *TC* method. We regard each detected temporal hotspot corresponds to one temporal interaction. From the result we can see that most of the interactions to the temporal hotspots are recalled; however, interactions without clear touching process in view are not able to retrieve. Table 2 shows the average probabilities of interaction detection for all experiences. All the hotspots show dominate interaction patterns with high probabilities (> 80%). This demonstrated that each hotspot has a single interaction pattern in this task, and all of them are proved to be correctly estimated by referring to the standard operation manual.

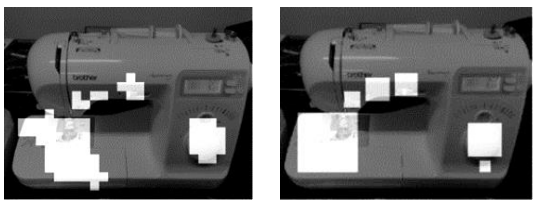

Figure 5.Comparison of hotspot resolution in one expert's experience. ***Fr*** in block level (left) and ***TC*** in connected-area level (right).

Table 1.Retrieve of temporal interactions

Participant	Interactions	Recall	Precision	F-Score
1	13345678	0.778	0.875	0.824
2	1345*679	0.778	1	0.875
3	134*5*679	0.778	1	0.875
4	1435*699	0.667	0.857	0.75
5	1345679	0.778	1	0.875

(*) is correct but not essential interaction. 1~9 standard interactions refer to the operation manual. 3 and 4 are changeable.

Table 2.Interaction patterns at hotspots.

Hotspot	Push	Put	Rotate	Slide
Spot 1	4%	**85.9**%	10.1%	0%
Spot 2	**92**%	0%	8%	0%
Spot 3	**90.8**%	0%	8%	1.2%
Spot 4	16.4%	0%	**83.6**%	0%
Spot 5	0%	3.6%	0%	**96.4**%

Conclusion

In this paper, we presented a novel method for detecting hotspots and interactions from recorded egocentric experiences. Our experimental result is encouraging and our pilot evaluation has demonstrated touch and hand motion, especially in depth, are good features for detecting hotspots and interactions. For potential applications of these hotspots, we can think of using them for: i) summarization of working experiences and providing guidance manuals, ii) working pattern analysis or machine usability analysis by interaction pattern analysis.

Overall, this work explores a novel area of capturing experiences with machines and how these can be described and used.

References

[1] Saba, Elliot N., et al. "Dante vision: In-air and touch gesture sensing for natural surface interaction with combined depth and thermal cameras." *ESPA*. 2012:167-170.

[2] Damen, Dima, et al. "You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video." *BMVC*. Vol. 2. 2014, 2: 4.

[3] Rogez, Grégory, et al. "Understanding Everyday Hands in Action from RGB-D Images." *ICCV*. 2015: 3889-3897.

[4] Agarwal, Ankur, et al. "High precision multi-touch sensing on surfaces using overhead cameras." *TABLETOP'07*. 197-200.

[5] Wilson, Andrew D. "Using a depth camera as a touch sensor." *ACM interactive tabletops and surfaces*. 2010: 69-72.

[6] Harrison, Chris, et al. "Omni-ouch: wearable multi-touch interaction everywhere." *24th ACM UIST*. 2011: 441-450.

[7] Lei, Jinna, et al. "Fine-grained kitchen activity recognition using rgb-d." *2012 ACM ubicomp*. 2012: 208-211.

[8] Leelasawassuk, Teesid, et al. "Estimating visual attention from a head mounted IMU." *ACM ISWC*. 2015: 147-150.

[9] Shimada, Nobutaka, et al. "Real-time 3D hand posture estimation based on 2D appearance retrieval using monocular camera." *ICCV*. 2001: 23-30.