ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 *2008. No. 8(88)*
ELEKTRONIKA IR ELEKTROTECHNIKA

T125 *AUTOMATION, ROBOTICS*
AUTOMATIZAVIMAS, ROBOTECHNIKA

Evaluation of Hand Pointing System Based on 3-D Computer Vision

P. Serafinavičius, S. Sajauskas
Department of Electronics Engineering, Kaunas University of Technology,
50 Studentų str., Kaunas, Lithuania, e-mail: paulius.se@gmail.com
G. Daunys
Department of Electronics, Šiauliai University,
141 Vilniaus str., Šiauliai, Lithuania, e-mail: g.daunys@tf.su.lt

Introduction

The current human-computer interfaces are inadequate to take the full advantages of computers. Some of the applications like smart rooms, virtual reality, household, industrial robots, mobile devices and others require a richer set of interaction modalities. Hand pointing permits humans to use their most versatile instrument, their hands, in more natural and effective ways than currently possible. While most gesture recognition devices are cumbersome and expensive, hand pointing with computer vision is more flexible. Definitely, it faces some difficulties due to the hand's complexity, lighting conditions, background artifacts, and user differences.

The person's hand pointing system for interactive environments has several desired criteria: it should operate in real-time, be robust to changing lighting conditions and background, and be able to detect and track user head and hand.

Vision based interfaces

A combination of range data, color data and face pattern recognition is used to track humans [1]. This system can track multiple users and locate their heads. The sensor fusion scheme is reported to work well, even in crowded environments, and with remarkable accuracy. However, the system requires three computers and dedicated hardware, training of the neural network, and tracks only head position.

C. Wren proposes a system that builds and tracks a blob-based model of the human body [2]. The model is then used to interact with virtual characters. This system is based on adaptive background subtraction. Its main limitations are that a static background is required and only a fixed camera can be used.

S. Grange describes the Human Oriented Tracking (HOT) library in [3]. It was developed as a tool for building vision-based interfaces. The design centers on a sensor-fusion based tracker that can efficiently detect, segment, and follow human features such as head, hands, etc. HOT is designed to provide the good performance using consumer-level computer hardware and cameras.

Our proposed vision based interface is able to detect user's head and hand and track them, providing the 3D coordinates of the head and hand in real time. It uses modified Viola-Jones method for head and hand detection. The fast KLT (Kanade-Lucas-Tomasi) features stereo tracking algorithm together with skin-color information and concentration of features was developed. The main advantages of the system are that it was realized using open source library OpenCV and uses inexpensive two USB web cameras. Therefore, it has a great flexibility in various kinds of applications. The difference between our work and the other systems is that we combine color and stereo vision to achieve better tracking. This paper presents the description of the algorithms and methods we used and the evaluation of our improved hand pointing system [4].

Methodology of user's head and hand detection

The main modality of our system is user's hand. In order to estimate user's hand pointing direction using stereo vision we need at least one good reference point. Other researchers of human-computer interaction have found that humans usually look at the same direction while pointing their hands at it. Therefore, detecting user's head and tracking its center point could be a good selection as the reference point. The details on how to detect head and hand by OpenCV library methods were described in our previous paper [4]. Here we will consider recent improvements of our system.

A flow chart diagram of the current system's simplified functionality is shown in Fig. 1. During the initialization the system allocates images' buffers and loads configuration file with the parameters necessary for detection and tracking. Then detection of hand is started. User can adjust the area on the image for more convenient place in order to detect hand. The hand is selected to be detected first because its detection is more complicated due to hand's non rigid shape. We adapted an object detection method proposed by Viola and Jones [5] and customized

by Kolch [6]. Objects are learned during a training phase with AdaBoost of features that compare grey-level intensity in rectangular image areas. The advantage is that it can be implemented with integral images. During the detection, a pre-computation step produces a 2-D brightness integral. The sum of pixel values in arbitrary rectangular areas can then be computed in constant time. Detection of hand of arbitrary scale runs with about 20 (10) frames per second on a 640x480 sized video stream on a 3 (1.6) GHz desktop (laptop) computer.

The initial hand pose should be a vertically oriented flat hand with closed fingers, parallel to camera's image plane. This posture is highly identifiable nature against background noise and is a fail-safe detection condition [7]. The recognition is executed in the part of the camera's field of view that corresponds to a natural reaching distance in front of the right shoulder. The original object detection method is very sensitive towards in-plane rotations. We used a trained detector for small rotations of the same hand posture, allowing the posture to be performed at angles from 0 to 15^{O} in order to increase the usability.

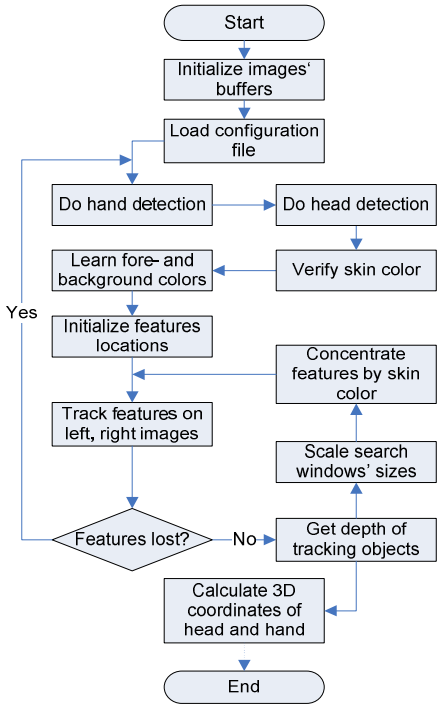

Fig. 1 Flow chart diagram of our proposed hand pointing system

Upon detection of the hand area, it is tested for the amount of skin colored pixels it contains. To this end, we used a histogram-based statistical model in HSV space from a large collection of hand-segmented pictures from many imaging sources, similar to Jones and Rehg's approach [8]. If a sufficient amount of area pixels are classified as skin pixels, the hand detection is considered successful and control is passed to the next stage, i.e. head detection. Similar, but simplified technique is used for head detection. While head detection is more reliable and faster than hand's, the skin color verification is unnecessary. During the development and evaluation we did not notice any head detection failures in our system. It is detected immediately after hand detection was verified. Only the condition is held: user's head must be in both cameras' fields of view and fit in it.

During the initialization of tracking the general statistical model of skin color is refined by learning the observed hand color on the detected area. This color histogram is contrasted to a reference area that is assumed to contain no skin areas, located around the hand area to the left, top and right. Other skin-colored objects, even other people that might be in this reference area are correctly considered background.

40 KLT features are placed on skin-colored spots with big eigen values in the detected area. In combination with image pyramids, a feature's image area can be matched efficiently to a similar area in the following video frame. KLT features do not encode object-level information. To achieve consistency among the features, to improve tracking across changing backgrounds, and to deal better with short occlusions, we enforce global constraints on the features' locations with a features concentration method that enforces the conditions of minimum and maximum feature distances.

In our approach, we used similar method to „Flocks of Features", described in [9]. At first the KLT features' positions are updated with the traditional pyramid-based feature matching algorithm. From their locations a small area is determined for further features tracking. Without additional effort, this would work fine only for rigid objects with a mostly invariant appearance. However, hands are highly articulated object whose appearance can change vastly and rapidly. The feature match correlation between two consecutive frames can thus be very low so that the feature must be considered as lost. Also, features might gradually move off the hand onto background areas with more prominent grey-level gradients. To cope with this situation and to better track the object at hand, our algorithm removes features with low correlation, those far from the centroid, and those too close to other features from the set. They are resurrected at good-to-track locations that also have a high skin color probability and are close to the cloud's centroid.

Tracking of corresponding features on right image is solved by the same KLT tracking method. The difference is that instead of previous frame image information we use current left cameras image and searching for features' matches in the right camera's image. Next, the refinement, i.e., concentration of features locations according to skin-colored pixels concentration in the right image is performed. This ensures reliability of stereo matching and avoids the usage of complex and slow disparity map calculation methods.

The calculation of head's and hand's 3-D world coordinates is performed after both left and right frame processing:

$$Z = b\frac{f}{x_L - x_R}; \quad (1)$$

$$X = x_L\frac{Z}{f}; \quad (2)$$

$$Y = y\frac{Z}{f}; \quad (3)$$

were b – a base distance, i.e., Euclidean distance between stereo cameras optical centers; f – focal length of stereo camera's lens; x_L, x_R, y – undistorted image coordinates [10].

The cameras must be calibrated according to methodology considered in [11]. Both, intrinsic and extrinsic stereo camera parameters must be provided to the system.

Experimental evaluation of the hand pointing system

The system must be evaluated by main aspects of computer vision methods, those are: speed, accuracy and robustness.

The speed of the system was measured by it self. We implemented real time frame rate calculation algorithm inside, which estimates frame rate, minimum and maximum latencies during both of the cameras' frames processing. The results are 15–23 frames per second (15–64 ms) on 3 GHz desktop computer and 6–15 frames per second (30–120ms) on 1.6 GHz laptop computer. Two Creative Webcam Live Pro USB cameras were used whose can supply up to 30 frames per second on 640x480 resolutions. While comparing with M. Kolsch's proposal, our system is also quite responsive, taking into account that we process frames from two cameras and track into them two objects (user's head and hand) in real-time. 300 ms is the threshold when interfaces start to feel sluggish and cause the "move and wait" symptom.

The methods of hand object detection were reused from M. Kolsch approach. As reported in [7], their methods were trained to have very low positive rate (<1e-10), thus achieving a very good detection rate of 85–95%. Head detection method was evaluated in our previous publication [4]. The result of evaluation was quite high, about 90%. Therefore, we can state, that our systems' object detection performance is acceptable for vision based user interface.

Evaluation of head and hand 3D tracking method robustness was also done. We define tracking robustness as ratio of successfully tracked hand to predefined targets with respect to all available targets. We had 3 users taking part in this experiment. Each of them was detected by the system (head and hand). Then user must track his hand to particular targets, marked on the images (Fig. 2), and return it back to initial position not loosing the feature points on a hand region in the cameras images. The points detected on user head region also should not be lost or reappear on other non-head regions while performing these tasks. If it happens, we evaluate this task as unsuccessful.

Fig. 2 The output windows of left and right cameras while evaluating tracking robustness of the hand pointing system

The results of tracking robustness while pointing hand to predefined eight targets are shown in the following tables. Each user made 3 tries to point the same target. There are successful tries for each target and the ratio R of them for each user.

Table 1. Lighting conditions: tungsten lamp, top-oriented

	R	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
U_1	0.96	2	3	3	3	3	3	3	3
U_2	0.92	2	3	2	3	3	3	3	3
U_3	0.71	1	2	1	3	3	2	3	2
			Avg.R		0.86				

Table 2. Lighting conditions: halogen lamp, front-oriented

	R	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
U_1	1.00	3	3	3	3	3	3	3	3
U_2	0.96	3	2	3	3	3	3	3	3
U_3	0.83	2	2	2	3	3	3	3	2
			Avg.R		0.93				

Table 3. Lighting conditions: tungsten lamp, twice higher intensity, top-oriented

	R	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
U_1	0.92	2	3	3	3	3	3	2	3
U_2	0.92	2	3	2	3	3	3	3	3
U_3	0.79	2	3	2	3	3	3	1	2
			Avg.R		0.88				

Table 4. Lighting conditions: natural lighting, front-oriented

	R	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
U_1	0.96	3	2	3	3	3	3	3	3
U_2	0.96	3	3	3	3	3	2	3	3
U_3	0.88	2	3	2	3	3	3	3	2
			Avg.R		0.93				

During the evaluation described above we also tested the accuracy in XY plane. Every user was able to point to the targets without extra effort. The failures were very occasional. But we still do not know about accuracy in depth (ZY plane). To ensure how accurate is our system while providing depth distances from camera to head or hand, we did the following experiment. User must be able to move in the perpendicular line away from cameras' plane. The line was graded with a step of 0,25m. The user moved slowly on a wheelchair from the possible closest distance to the far away until the end of the line, every 0,25m fixating his hand coordinates. The fixation was done manually by other person. User was asked to keep his hand in fixed state as it was possible and move only by the wheelchair along the defined line. We tested 8 times with the same user. The results how user hand's Z coordinate is related to real world distance are shown in Fig. 3.

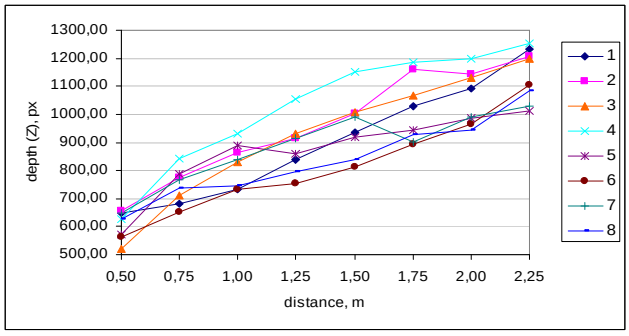

Fig. 3. Depth accuracy evaluation. Hand's Z coordinate in relation to real world distance. Eight tries were performed with one user

Conclusion

The evaluation of the hand pointing system was performed by the main aspects of computer vision methods: speed, accuracy and robustness.

The speed (responsiveness) of the system was 15–23 fps (15–64ms) on 3 GHz desktop computer and 6–15 fps (30–120ms) on 1,6 GHz laptop computer. It is acceptable for vision based user interfaces.

Detection rate of user head and hand is quite high while using our detection algorithm based on the methodology, proposed by Viola, Jones [5] and Kolch [7]. The average rate of correctly detected head and hand regions is 90%.

Average robustness rate of tracking algorithm now is 90% instead of 58%, evaluated on the previous system of ours [4]. Variance of robustness rate on lighting conditions was small. However, this approach requires good lighting conditions to ensure high robustness in general.

The depth information provided by the system meets the real world distances. Therefore, accuracy of the system is acceptable for 3D user interface.

References

1. **Darrell T., Gordon G., Harville M., Woodfill J.** Integrated person tracking using stereo, color, and pattern detection // Proceedings of the Conference on CVPR. – Santa Barbara. – 1998. – P. 601–609.
2. **Wren C., Azarbayejani A., Darrell T., Pentland A.** Pfinder: Real-Time Tracking of the Human Body // IEEE Transaction on Pattern Analysis and Machine Intelligence. – 1997. – Vol. 19, No. 7. – P. 780–785.
3. **Grange S., Casanova E., Fong T., Baur C.** Vision-based sensor fusion for Human-Computer Interaction // IEEE/RSJ International Conference on Intelligent Robots and Systems. – Lausanne. – 2002.
4. **Serafinavičius P.** Estimating Characteristic Points of Human Body for Automatic Hand Pointing Gesture Recognition // Electronics and Electrical Engineering. – Kaunas: Technologija, 2007. – No. 8(80). – P. 83–86.
5. **Viola P., Jones M.** Robust Real-time Object Detection // International Workshop on Statistical and Computational Theories of Vision. – 2001.
6. **Kolsch M., Turk M.** Analysis of Rotational Robustness of Hand Detection with a Viola-Jones Detector // IAPR International Conference on Pattern Recognition. – 2004.
7. **Kolsch M., Turk M.** Robust Hand Detection // Proc. IEEE International Conference on Automatic Face and Gesture Recognition. – 2004.
8. **Jones M. J., Rehg J. M.** Statistical Color Models with Application to Skin Detection // International Journal of Computer Vision. – 2002. – No. 46(1). – P. 81–96.
9. **Kolsch M., Turk M.** Fast 2D Hand Tracking with Flocks of Features and Multi-Cue Integration // Computer Vision and Pattern Recognition Workshop. – 2004. – Vol., Issue 27–02. – P. 158–158.
10. **Serafinavičius P., Daunys G.** Detection of Hand Position using 3-D Computer Vision // Electronics and Electrical Engineering. – Kaunas: Technologija, 2006. – No. 7(71). – P. 63–66.
11. **Serafinavičius P.** Investigation of Technical Equipments in Computer Stereo Vision: Camera Calibration Techniques // Electronics and Electrical Engineering. – Kaunas: Technologija, 2005. – No. 3(59). – P. 24–27.

Received 2008 03 31

P. Serafinavičius, S. Sajauskas, G. Daunys. Evaluation of Hand Pointing System Based on 3-D Computer Vision // Electronics and Electrical Engineering. – Kaunas: Technologija, 2008. – No. 8(88). – P. 95–98.

Evaluation of hand pointing system is presented. The system is based on 3D computer vision and implemented using open source computer vision library (OpenCV). The system is able to detect and track user's head and hand and return 3D coordinates in real time. The detection of head and hand is based on Viola-Jones detector applying human skin color model information. The detection rate is about 90%. The tracking is based on Kanade-Lucas-Tomasi iterative algorithm, customized for 3D computer vision case. High tracking robustness (90%) is achieved due to skin color model and variable size search window according to depth information. The speed of the system is 15–23 frames per second on 3 GHz desktop PC. Ill. 3, bibl. 11 (in English; summaries in English, Russian and Lithuanian).

П. Серафинавичюс, С. Саяускас, Г. Даунис. Оценка опознавательной системы показательных движений руки на базе компьютерного стереозрения // Электроника и электротехника. – Каунас: Технология, 2008. – № 8(88). – С. 95–98.

Представлена экспериментальная оценка системы показательных движений руки. Эта система была создана на базе стереозрения и реализована при помощи библиотеки компьютерного зрения открытого кода (OpenCV). Система способна обнаружить и следить за движением головы и руки пользователя в течение реального времени и вернуть их трехмерные координаты. Детектор головы и руки был изобретен на основе метода детектирования Виола-Джонс, используя информацию модели цвета человеческой кожи. Надежность детектирования достигает 90 %. Основой следования головы и руки является итеративный алгорифм Канаде-Лукас-Томаси, приспособленный для стереозрения. Высокая надежность следования (90 %) достигается при помощи модели цвета кожи и меняющейся величины поискового окна в зависимости от информации глубины. Скорость системы 15–23 кадра в секунду на 3 ГГц стационарном компьютере. Ил. 3, библ. 11 (на английском языке; рефераты на английском, русском и литовском яз.).

P. Serafinavičius, S. Sajauskas, G. Daunys. Rodomųjų rankos judesių atpažinimo sistemos, pagrįstos kompiuterine stereorega, įvertinimas // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2008. – Nr. 8(88). – P. 95–98.

Pateikiamas rodomųjų rankos judesių atpažinimo sistemos eksperimentinis įvertinimas. Ši sistema buvo sukurta stereoregos pagrindu ir įgyvendinta panaudojus atvirojo kodo kompiuterinės regos biblioteką (OpenCV). Ji gali aptikti ir sekti žmogaus galvos ir rankos judesius realiu laiku bei grąžinti jų trimates koordinates. Galvos ir rankos judesių detektorius buvo sukurtas Viola-Jones detektavimo metodo pagrindu, naudojant žmogaus odos spalvos modelio informaciją. Detektavimo patikimumas siekia 90 %. Rankos ir galvos sekimo pagrindas yra Kanade-Lucas-Tomasi iteracinis algoritmas, pritaikytas stereoregos atvejui. Didelis sekimo patikimumas (90 %) gaunamas odos spalvos modeliu, įvertinant pagal gylio informaciją kintančio paieškos lango dydį. Sistemos greitaveika, naudojant 3 GHz stacionarų kompiuterį, yra 15–23 kadrai per sekundę. Il. 3, bibl. 11 (anglų kalba; santraukos anglų, rusų ir lietuvių k.).

DOI: 10.5755/j02.eie.11359