REAL-TIME MULTI-CANDIDATES FUSION BASED HEAD TRACKING KINECT DEPTH SEQUENCE

Zhiting Yang[1], Yang Yang[1], Yun-Xia Liu[2]

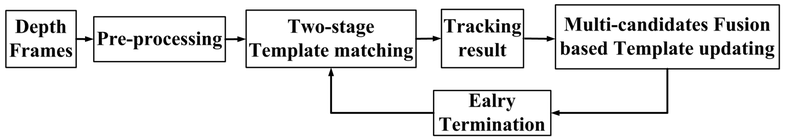
[1]School of Information Science and Engineering, Shandong University, Jinan, China

[2]School of Control Science and Engineering, Shandong University, Jinan, China

Abstract

Considering depth images are robust to illumination variations with complex backgrounds, the paper developed a real-time head tracking system with one Kinect camera. Distance transform is applied to pre-processed depth frames to further reduce the effect of appearance deformation. A multi-candidates fusion strategy is proposed for template updating that assures head representation robustness. Two-stage template matching is adopted for computational efficiency in the searching procedure. In addition, an early termination criterion for template updating is presented to reliably improve the tracking accuracy. Abundant experimental results on our depth database demonstrate that the proposed method performs favorably against state-of-the-art methods in terms of robustness, accuracy, and efficiency.

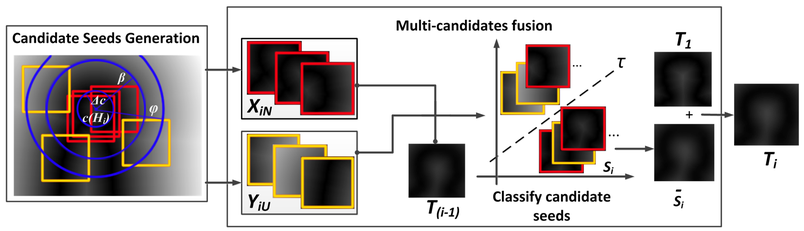
Flowchart of the proposed method



Pre-processing

A series of morphological operations including complement, erosion and dilation are adopted to fill in occlusion regions. To further reduce the effect of appearance deformation, we adopt distance transform to solve the problem of low level of discrimination of depth images and enrich detail information.

Flowchart of multi-candidates fusion



Candidate seeds generation

A bunch of depth image patches are selected as ***candidate seeds*** for template updating within a circular area centered at $c(H_i)$, where H_i denotes the depth patch of tracking head in the i-th frame, and $c(\bullet)$ denotes their center location. Our previous work reveals that the number of pixels corresponding to each of 10 centimeters (decimeter, dm) under different distances conditions was modeled as

$$\alpha = \frac{5.11\times 10^4}{d_s - 56.5}$$

where $d_s = (255 - D_i)\times 16$ represents the physical distance d_s between head and camera, and the depth data D_i can be extracted easily on depth images. Let Δc denote the number of pixels that the head location changes between adjacent images at most. With reasonable assumption that a common person's average walking pace is 1m/s (10 dm/s) and the Kinect camera captures 30 images per second, we have

$$\Delta c = \frac{10}{30}\alpha = \frac{1.7\times 10^4}{d_s - 56.5}$$

Taking into account abruptly rapid motion of the tracking object, we also allow candidate seeds to be selected out of the radius . The set of candidate seeds can be randomly selected as

$$\begin{cases} X_{iN} = \left\{ X_{ij} \middle| \left| c(X_{ij}) - c(H_i) \right| \le \Delta c,\ j = [1,\cdots,N] \right\} \\ Y_{iU} = \left\{ Y_{it} \middle| \beta \le \left| c(X_{ij}) - c(H_i) \right| \le \varphi, t = [1,\cdots,U] \right\} \end{cases}$$

where β and φ are inner and outer radius ($\Delta c \le \beta \le \varphi$) of the ring area, N and U are constant parameters that denote the cardinality of X_{iN} and Y_{iU}.

Multi-candidates fusion based template updating

Patches extracted from candidate seed positions that are not related to the object to be tracked should be left out during template updating. Only those from related seed positions are utilized as ***candidates*** for fusion to generate the updated template for the next frame.

The proposed candidate determination algorithm can be formulated as following. For each candidate seed , we calculate their correlation with the template of last frameT_{i-1},

$$corr_{ip} = \frac{\sum_m \sum_n \left(\left(c_{ip\,mn} - \overline{c_{ip}} \right) \left(T_{i-1mn} - \overline{T_{i-1}} \right) \right)}{\sqrt{\left(\sum_m \sum_n \left(c_{ip\,mn} - \overline{c_{ip}} \right)^2 \right) \left(\sum_m \sum_n \left(T_{i-1mn} - \overline{T_{i-1}} \right)^2 \right)}}, T_0 = H_1$$

where m and n are the row and column indexes of image patches, $\overline{c_{ip}}$ and $\overline{T_{i-1}}$ are means of c_{ip} and T_{i-1} ,respectively. Then we get the set of candidates for the i-th frame as

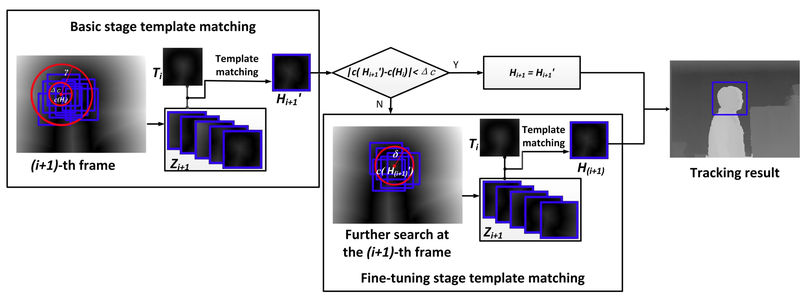
$$S_i = \left\{ c_{ip} \middle| corr_{ip} \ge \tau \right\}$$

where τ is a threshold parameter. Then the multi-candidate fusion based template T_i at the i-th frame is derived by

$$T_i = \rho_1 \times \overline{S_i} + \rho_2 \times H_1$$

where $\overline{S_i}$ is the mean of candidate sets, ρ_1, ρ_2 are weighting coefficients that satisfies . Template T_i is used for matching in the *(i+1)*-th frame to give the tracking result.

Flowchart of the two-stage template matching



Two-stage template matching strategy

A two-stage template matching strategy is proposed to address the contradiction between tracking effectiveness and computation efficiency. Different sampling ratios are employed for nearer and further test samples, i.e.

$$\Delta s = \begin{cases} 1, & \text{if } \left| c(Z_{(i+1)k}) - c(H_i) \right| \le \Delta c \\ 1/\delta, & \text{if } \Delta c \le \left| c(Z_{(i+1)k}) - c(H_i) \right| \le \gamma \end{cases}$$

where γ denotes the search range and $Z_{(i+1)k}$ represents the k-th test sample in the *(i+1)*-th frame, $k = [1,\cdots,V]$ is the index for test samples. According to this sampling rule, all test samples Z_{i+1} could be extracted and correlation

$$corr_{(i+1)k} = \frac{\sum_m \sum_n \left(\left(Z_{(i+1)k\,mn} - \overline{Z_{(i+1)k}} \right) \left(T_{imn} - \overline{T_i} \right) \right)}{\sqrt{\left(\sum_m \sum_n \left(Z_{(i+1)k\,mn} - \overline{Z_{(i+1)k}} \right)^2 \right) \left(\sum_m \sum_n \left(T_{imn} - \overline{T_i} \right)^2 \right)}}$$

could be computed. A *temporary* head location $c(H_{i+1}')$ that maximize the correlation coefficients could be obtained:

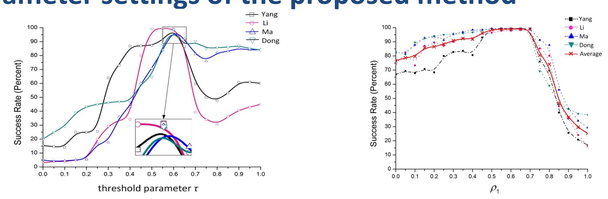
$$c(H_{i+1}') = \arg\max_k corr_{(i+1)k}$$

We also record the maximum correlation coefficient as $corr_{max}$. This forms the ***basic stage*** of template matching. A second ***fine-tuning stage*** is designed to enhance the tracking accuracy. We carry out exhaustive search centered at $c(H_{i+1}')$ within radius to obtain the more accurate head location $c(H_{i+1})$, and $corr_{max}$ should also be updated. Note that the second fine-tuning stage is optional and only take place when $\left| c(H_{i+1}') - c(H_i) \right| \ge \Delta c$.

Early termination criterion for template updating

In case in the *(i+1)*-th frame, we can conclude that T_{i+1} is not similar to T_i. To prevent tracking drift, we stop template updating by setting T_{i+1}=T_i and expanding γ to $\gamma*$ ($\gamma < \gamma*$) at the *(i+2)*-th frame.

Parameter settings of the proposed method



It is good property to be robust to parameter settings. In this subsection, we discuss the settings of τ and ρ_1 in an experimental manner. In our experiments, τ is fixed to 0.6 and ρ_1 is fixed to 0.7.

Subjective comparison of tracking results

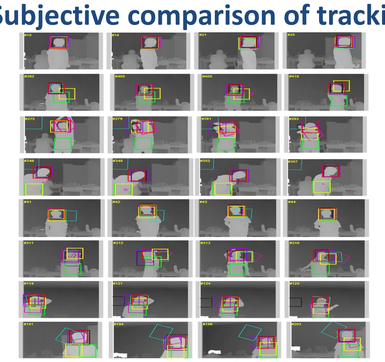


Figure shows part of tracking results of the proposed method under the varying circumstance of motion, self-occlusion, deformations and camera motion from top to bottom. It is safely concluded that our method is highly robust and adaptable. However others are not suitable for depth image tracking. As clearly depicted in Table that the proposed method yields the best tracking result. An average 99% success rate is achieved, which demonstrates its robustness. On the other hand, some tracker that works well in color channels loses their superiority (below 50%) when directly applied to depth sequences.

Sequence	Frame number	Proposed	TLD [2]	SemiB [17]	SCM [18]	CT [19]	IVT [20]	FCT [3]
Yang	500	**100**	92	84	29	30	56	7
Ma	455	**98**	81	48	37	21	19	17
Dong	422	**98**	90	39	72	47	47	31
Li	468	**99**	96	84	96	53	35	34
Zhang	431	**100**	87	91	74	72	72	52
Kong	429	**99**	82	94	21	53	24	15
Average		**99**	88	73	54	46	42	26