早稲田大学大学院情報生産システム研究科

博 士 論 文 概 要

論　文　題　目

Temporal Mapping and Temporal Prediction based Ultra-low Delay Object Tracking Computing Architecture for Visual Feedback System

申　請　者

Tingting HU

情報生産システム工学専攻
画像情報システム研究

２０２３年　４月

Visual feedback system, which combines virtual (computing) and real spaces by sensing the status of real space and feedbacking valued information to real space after visual processing in virtual space, draws increasing attention in fields such as robotics, factory automation (FA), and entertainment. Object tracking, which includes target detection and tracking, is one of the most important visual processing techniques for real-time applications. On the other hand, the change in real space continues while visual processing is running in virtual space, making the ultra-low delay critical to highly integrate real and virtual spaces. Among various visual processing architectures, FPGA-based stream architecture is the one that is possible to achieve both complex spatial processing and delay lower than 1 ms while suitable algorithms and architectures are needed. Therefore, the FPGA-based ultra-low delay object tracking algorithm and architecture become important.

FPGA-based stream architecture achieves ultra-low delay by performing visual processing while transmitting the image data from the sensor to the processor in a fully pipelined way. The complex spatial processing in the object tracking algorithms has always been a problem, preventing object tracking from being implemented with fully pipelined architecture. Here introduces three representative complex spatial processing. Spatial multi-range search processing, which performs the same processing for different parameters and selects the result with the most suitable parameter, exists in target detection. Lots of resources are required due to the extensive processing performed in the spatial domain for wide-range parameter search. Dependent sequential processing, which performs two dependent processing in a sequential way, exists in target tracking. The dependent relation for two processing performed in a local spatial area generates not only high resource costs but also high delay. Spatial iterative processing, which performs the same processing iteratively with the output of the previous processing as the input, exists in target tracking. Iterative processing of a large amount of spatial processing not only consumes expensive resources but also causes a significant delay. On the other hand, the difference between the output state from virtual space and the true state in the real space caused by delay remains, albeit greatly reduced by ultra-low delay processing. Especially for target tracking, the difference introduces tracking errors.

To solve the above problems and achieve ultra-low delay object tracking architecture, temporal mapping and temporal prediction are proposed based on high temporal resolution characteristics of the ultra-low delay system. Temporal mapping takes advantage of the small difference

between successive frames of a sequence taken at the high temporal resolution, simplifying and reducing the amount of processing in the spatial domain by mapping a large amount of processing performed in the spatial domain to the temporal domain. Temporal prediction takes advantage of the detailed moving information captured at the high temporal resolution, compensating for changes due to movement by predicting changes in the temporal domain using a prediction model. Combining these two methods, the aforementioned problems are solved.

This dissertation is organized as follows.

In Chapter 1, the background of the dissertation is described, including the widespread applications of visual feedback systems, the importance of object tracking, the importance of the ultra-low delay, the problems, and the concept of proposals. Furthermore, the target and organization of this dissertation are shown.

In Chapter 2, aiming at an ultra-low delay target detection, temporal template prediction-based keypoint matching is proposed. Keypoint matching-based target detection, which consists of feature detection and feature matching, is one of the popular target detection algorithms. Handling scale change is one of the most challenging tasks. The conventional method [IEICE, 2013] proposed the multi-templates method, which handles size change by preparing multiple templates with various sizes for feature matching and searches for the template with the best matches. With a large-scale of spatial processing performed in the spatial domain, lots of resources are needed for FPGA implementation. Given that the scale change occurs continuously in the visual system with the high temporal resolution, it becomes possible to predict the size of the template for the next frame using the current existing matching results. The proposed method maps the feature-matching processing for various sizes of templates into the temporal domain based on the temporal template prediction rule, significantly reducing the complexity that exists in the feature matching. The proposed method is implemented as a practical system by integrating a high-speed camera and an FPGA. Hardware evaluation shows that the proposed method is capable of handling a wide range of scale changes (11 templates) at a low resource cost (<80%). Additionally, it also shows that the designed target detection system is capable of sensing and processing 1000 fps sequence (Resolution: 640×360) with a delay of less than 1 ms/frame.

In Chapter 3, to achieve ultra-low delay target tracking with high real-time accuracy, temporal prediction-based parallel motion estimation and temporal iterative tracking are proposed. Differential-based target

tracking estimates the motion between two image patches based on derivatives of image intensities. Estimating both translation and rotation changes is crucial in the practical applications. The conventional method [CVIM, 2003] estimates the translation first, and rotation is estimated using the image patch that has been warped in accordance with the estimated translation. The data dependency-based sequential processing introduces intermediate processing, leading to significant delays in addition to resource expenditures. The proposed method breaks the data dependency between translation and rotation estimations by mapping the data dependency into temporal domain based on temporal prediction, allowing the translation and rotation estimations performed in a parallel way. The independent processing style avoids the problems caused by intermediate processing. High accuracy over the subpixel level is usually required for precise location. The conventional method [CVPR, 1994] applies Newton-Raphson iteration to direct a more accurate motion estimation. However, the large number of iterative processing performed on the spatial domain results in significant resource costs in addition to delay. On the other hand, the target continues moving while the tracking processing is running, making the tracking error depend not only on the image processing itself but also on the delay. The proposed method maps the spatial iterative motion estimation into the temporal domain to avoid the large number of spatial processing, and predicts the motion that occurs during the tracking processing ongoing to reduce the tracking error caused by delay. The proposed methods are implemented as a practical system by integrating a high-speed camera and an FPGA. Algorithm evaluation shows that the proposed method achieves subpixel level real-time accuracy, and outperforms other related methods. Hardware evaluation shows that the designed target tracking system supports sensing and processing 1000 fps sequence (Resolution: 640×360) with a delay of less than 1 ms/frame, and costing resources less than 20%.

In Chapter 4, the overall dissertation is summarized, and the future works are described. To realize ultra-low delay target detection and target tracking, temporal mapping and temporal prediction based algorithms and architectures are proposed. Both proposed target tracking and detection systems achieve delays of less than 1 ms/frame when processing a 1000 fps sequence with a resolution of 640×360.