Accuracy Evaluation of Second-Order Shape Prediction on Tracking Non-Rigid Objects

Kenji Nishida
and Takumi Kobayashi

National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, JAPAN
Email: kenji.nishida@aist.go.jp takumi.kobayashi@aist.go.jp

Jun Fujiki

Department of Applied Mathematics,
Fukuoka University
Fukuoka, JAPAN
Email: fujiki@fukuoka-u.ac.jp

***Abstract*—For our previously proposed shape prediction based tracking algorithm for non-rigid objects, the shape prediction accuracy is critical for the tracking performance. Therefore, we have presented a preliminary evaluation of second-order shape prediction algorithm for tracking non-rigid objects. In the proposed algorithm, the object shape was predicted from the movement of feature points, which were approximated by a second-order Taylor expansion. Approximate first-order movements, the so-called optical flows, were simultaneously exploited by chamfer matching of edgelets. Shape prediction accuracy was evaluated by chamfer matching between the predicted object shape and the actual object shape. While only one video sequence was preliminary evaluated, three more video sequences were evaluated. The new sequences are captured by fixed camera, while our previous sequence was captured by Hand-held camera. In this paper, the effect of second-order shape prediction is quantitatively analyzed by more video sequences. The method exhibits superior shape prediction performance compared to a simple linear prediction method.**

Keywords–Tracking non-rigid objects; Chamfer distance; Shape prediction; Optical flow.

I. INTRODUCTION

Visual object tracking is one of the most popular techniques in the field of computer vision. We have proposed a novel algorithm for tracking non-rigid (deformable) objects based on the second order shape prediction and presented a preliminary evaluation of its performance against the linear (first-order) prediction measured by the similarity between the predicted and actual shapes of the tracked object [1].

Recently, tracking algorithms for non-rigid (deformable) objects have been used in many application fields [2], [3]. In sports scenes, especially those of team sports such as football, there are many objects in similar appearance, which increase the difficulty of tracking. Therefore, we consider both the movement and shape (form) of these objects to be discriminative for tracking.

For the shape of the non-rigid object to change in every video frame, next object shape have to be predicted from preceeding video frames to identify the object. A number of human pose estimation algorithm has been proposed, such as 3D pose estimation of an articulated body using template matching [4] and matching algorithm of pictorial structures [5]. However, they are not predicting the pose in the next video frames.

The movement of the parts must be detected to predict the object shape, and the smallest part must be a feature point. When the object shape is represented by the collection of feature points, the deformation of the object is predicted by exploiting the movement of the feature points. Sim and Sundaraj proposed a motion tracking algorithm using optical flow [6], and this can be considered as the first-order approximation of the movement. For our tracking algorithm, we adopted a shape prediction algorithm based on the second-order approximation of the feature points' movement [7].

In this paper, the evaluation was applied to three more video sequences, which were captured by a fixed camera; Tai chi chuan demonstration, a skier's backshot, and a skier's frontshot. They were compared with the previously evaluated sequence of a skier, which was captured by a hand-held camera. Thereby, the effect of object movement and camera ego-motion were examined from these results.

The remainder of this paper is organized as follows. In Section II, we summarize the previous tracking algorithms. In Section III, we describe our shape prediction algorithm and the tracking procedure that uses the chamfer distance as a similarity measure. The experimental results are presented in Section IV. Finally, we present our conclusions and ideas for future work in Section V.

II. PREVIOUS TRACKING ALGORITHMS

The primary function of an object tracking is to find a moving object in an image. Therefore, detecting differences between consecutive video frames adopted in the first approach, such as a background subtraction algorithm which was employed by Koller [8]. However, the static background might be required, and obviously object detection was difficult when the movement of the objects was small,

A group of feature-based tracking algorithms [9], [10], [11] is proposed as the second approach. Salient features such as corner features are individually extracted and tracked are grouped as belonging to the corresponding object. It can be robust to illumination change. However, the precision of the object location and dimension is affected by the difficulties that arise in feature grouping. The mean-shift algorithm [12], [13] is also included in the feature-based algorithms. In mean-shift algorithm, the local features (such as color histograms) of pixels belonging to the object are followed. The mean-shift approach allows robust and high-speed object tracking, if a local feature that successfully discriminates the object from

the background exists. However, it is difficult to discriminate objects that are close to each other and are similar in color, or to adopt this method for gray-scale images.

Avidan redefined the tracking problem as that of classifying (or discriminating between) the objects and the background [14]. This third approach can be categorized as a detect-and-track approach. In this approach, features are extracted from both the objects and the background; then, a classifier is trained to classify (discriminate) the object from the background. Grabner trained a classifier to discriminate an image patch within an object in the correct position and those with objects in the incorrect position [15], and thereby, the position of the object could be estimated more precisely. While this approach allows stable and robust object tracking, a large number of computations are necessary. The approach of Collins and Mahadevan is regarded as an approach of this type, but they selected discriminative features instead of training classifiers [16], [17]. Grabner introduced on-line boosting to update feature weights to attain compatibility between the adaptation and stability for the appearance change (illumination change, deformation, etc.) of tracking classifiers [18]. Woodley employed discriminative feature selection using a local generative model to cope with appearance change while maintaining the proximity to a static appearance model [19]. The tracking algorithms are also applied to the non-rigid (deforming) objects. Godec proposed *Hough-based tracking* algorithm for non-rigid objects, which employed Hough voting to determine the object's position in the next frame [3].

In detect-and-track approaches, the estimated object position in the next video frame is determined based on the similarity of the features to the object in the current video frame, and a change in appearance, especially deformation, may affect the similarity between the object in the current and the next frame, and thereby, the accuracy of the tracking. Therefore, the tracking accuracy can be improved by predicting the deformation of the object to improve the similarity of the object in the next video frame to that in the current video frame. Sundaramoorthi proposed a new geometric metric for the space of closed curves, and applied it to the tracking of deforming objects [2]. In this algorithm, the deforming shapes of the objects are predicted from the movement of the feature points using first order approximation. However, the first-order approximation is not sufficient to estimate the reciprocating movement, which often human legs and arms do.

III. Shape-based Predict-and-Track Algorithm

In this section, we describe an algorithm for tracking by shape prediction [7]. The algorithm consists of two components, shape prediction and tracking by shape similarity.

A. Notation

The following notation is used throughout this paper.

- X denotes the center of the object,
- $O(X)$ denotes the object image centered at position X,
- $E(X)$ denotes the binary edge image for the object at position X,
- $\hat{O}$ and $\hat{E}$ denote the predicted image and edge image of the object, respectively,
- x denotes the positions of the feature points for object X,
- x' denotes the differential of x, i.e., $x' = \frac{dx}{dt}$,
- x'' denotes $\frac{d^2x}{dt^2}$,
- $\tilde{x}$ denotes the subset of feature points in the object that constitute the outline edge, $\tilde{x} \in E(X)$,
- $\hat{x}$ denotes the predicted position at the next frame for $\tilde{x}$,
- $l(x)$ denotes the edgelet for position x.

B. Shape Prediction

The object shape is represented by the collection of feature points x, and the deformation of the object is predicted by exploiting the movement of the feature points.

Let x_t be the 2-D position of the feature points that constitute the object image O at time t. The position of the points at $t+1$ can be estimated using a Taylor expansion. Up to the second-order, this is

$$x_{t+1} = x_t + x'_t + \frac{1}{2}x''_t, \tag{1}$$

where x' is the so-called optical flow, which is practically computed as the difference in the pixel position:

$$x'_t = x_t - x_{t-1}. \tag{2}$$

Similarly, x'' denotes the second-order differential of x, which is calculated as

$$\begin{aligned} x''_t &= x'_t - x'_{t-1} \\ &= x_t - x_{t-1} - (x_{t-1} - x_{t-2}) \\ &= x_t - 2x_{t-1} + x_{t-2}. \end{aligned} \tag{3}$$

Therefore, the appearance of the object at $t+1$ can be predicted based on the feature point movements computed from three consecutive video frames. Suppose that the shape of the object is determined by the outline edge image Es. The algorithm for detecting the feature point movements is described in Section II-D.

C. Estimation of Object Translation

The movement of the feature points comprises both the object translation (global movement of the center of the object) and the movement relative to the center of the object, which is described by

$$x'_t = X'_t + r'_t, \tag{4}$$

where X denotes the position of the object's center, and r denotes the position of the pixels relative to X. Figure 1 shows the movement of feature point x', the movement of the object's center X', and the relative movement r'.

The relative movement r' is derived from the object deformation, and thus makes a significant contribution to the prediction of the object's shape. Because relative movement obeys the physical constraints of the body parts of the object, its second-order prediction is effective. In contrast, the second-order movement contributes less to the object translation X, because such global movement is considered to be smooth ($X' \approx 0$). Therefore, the purpose of our tracking algorithm is to determine the next object position X_{t+1} based on the similarity between the predicted and actual object shapes, which is computed globally.

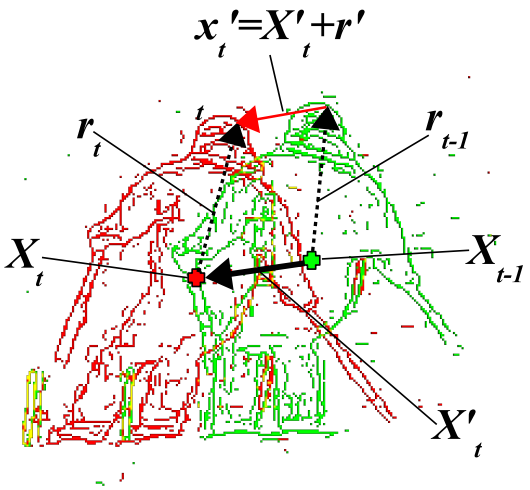


Figure 1. Edge image and object movement.
Green: Edge image for $t-1$, Red: Edge image for t

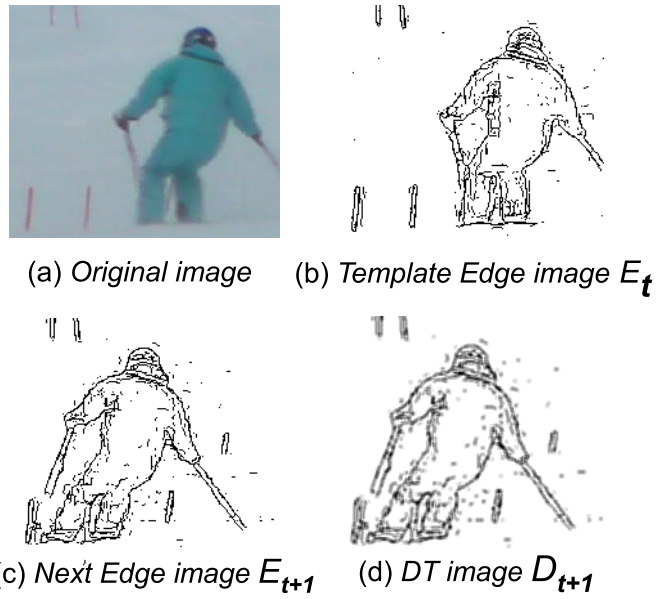


Figure 2. Chamfer system.

The similarity between the predicted edge image $\hat{E}_{t+1}$ and actual edge image E_{t+1} is measured using the chamfer system [20]. This system measures the similarity of two edge images using a distance transform (DT) methodology [21].

Let us consider the problem of measuring the similarity between template edge image E_t (Figure 2(b)) and a successive edge image E_{t+1} (Figure 2(c)). We apply the DT to obtain an image D_{t+1} (Figure 2(d)), in which each pixel value d_{t+1} denotes the distance to the nearest feature pixel in E_{t+1}. The chamfer distance $D_{chamfer}$ is defined as

$$D_{chamfer}(E_t, E_{t+1}) = \frac{1}{|E_t|} \sum_{e \in E_t} d_{t+1}(e), \quad (5)$$

where $|E_t|$ denotes the number of feature points in E_t and e denotes a feature point of E_t.

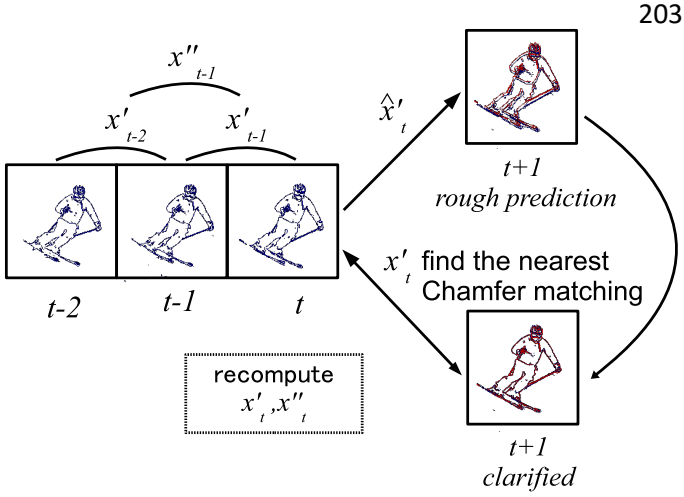


Figure 3. Tracking procedure.

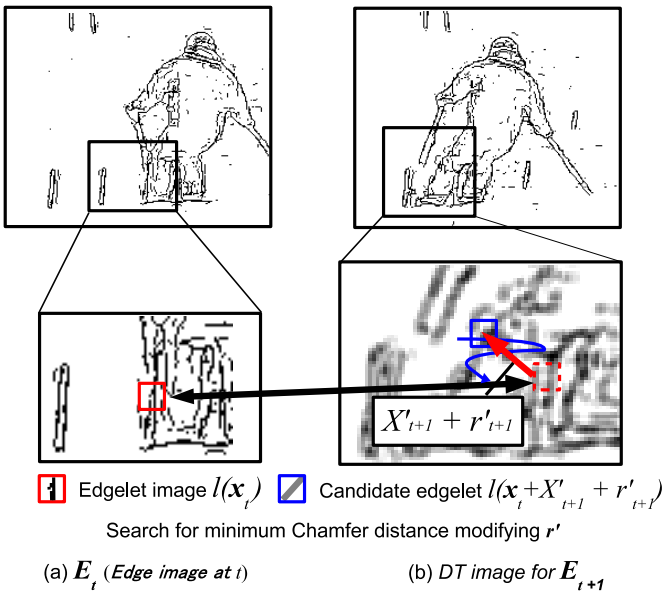


Figure 4. Edgelet tracker.

The translation of the object can be estimated by finding the position of the predicted edge image $\hat{E}_{t+1}$ that minimizes $D_{chamfer}$ between $\hat{E}_{t+1}$ and the actual edge image E_{t+1}:

$$X_{t+1} = \arg\min_{E_{t+1}} D_{chamfer}(\hat{E}_{t+1}, E_{t+1}). \quad (6)$$

Figure 3 illustrates the tracking procedure. First, the optical flow x'_t and its approximate derivative x''_t are computed from preceding video frames at $t-2$, $t-1$, and t. The object shape at $t+1$, denoted by $\hat{E}_{t+1}$, is then predicted using x' and x''. The object position is determined by locating $\hat{E}_{t+1}$ at the position of minimum chamfer distance to the actual shape at $t+1$, E_{t+1}. Finally, the optical flow for the next video frame x'_{t+1} is recomputed using actual edge images E_t and E_{t+1}.

D. Detection of Feature Point Movements

After the object translation X'_{t+1} has been determined, the movement of the feature points x'_{t+1} is detected from the actual object images $O(X_t)$ and $O(X_{t+1})$.

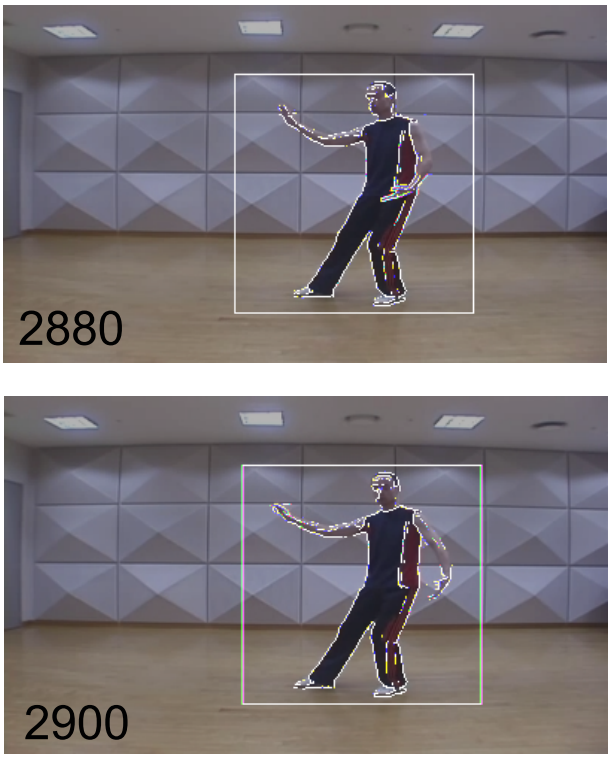


Figure 5. Tracking and shape prediction for Tai chi chuan.
Blue: Ground Truth; Green: Linear Prediction; Red: Second-order Prediction.

The feature point movements x'_{t+1} are directly computed based on the actual edge image at $t+1$ by tracking small parts of the edge (edgelets). We also employed the chamfer system to detect the movement of the edgelets. A template edgelet image $l(\tilde{x}_t)$ extracted from E_t is compared against the candidate edgelet $l(\tilde{x}_t + \hat{x}'_{t+1})$ in the next edge image E_{t+1}. By minimizing the chamfer distance between the two, we obtain the feature point movement (Figure 4):

$$\hat{x}'_{t+1} = \arg\min_{\hat{x}'_{t+1}} D_{chamfer}(l(\tilde{x}_t), l(\tilde{x}_t + \hat{x}'_{t+1})). \quad (7)$$

As the detected movements $\hat{x}'_{t+1}$ may contain noise, we apply a smoothing process by averaging the relative movements in the neighboring region:

$$x'_{t+1} = \frac{1}{N} \sum_{\hat{x}'_{t+1} \in \delta_{t+1}} \hat{x}'_{t+1}, \quad (8)$$

where N denotes the number of detected movements $\hat{x}'_{t+1}$ in the neighborhood δ of $\tilde{x}_t$.

IV. Evaluation of Shape Prediction Performance

The algorithm described above was applied to three video sequences (Tai chi chuan demonstration, a skier backshot, and a skier frontshot) captured by fixed camera, and a sequence of a skier captured by a hand-held camera. The effect of object translation and camera ego-motion on the shape prediction performance was examined.

Figure 6. Shape prediction accuracy for Tai chi chuen.

The proposed second-order shape prediction model was compared with linear one, which is formulated by

$$x_{t+1} = x_t + x'_t. \quad (9)$$

The prediction performance was evaluated by the chamfer distance between the actual image E_{t+1} and the predicted image $\hat{E}_{t+1}$ using (5).

A. Video sequences captured by fixed camera

1) Tai chi chuan demonstration: Figure 7 shows the video frames 3006–3009, when the linear prediction attained better precision. The both prediction algorithm had large error in the shape of right knee in the frame 3006, because of the quick motion by the player. The error in the estimation of feature point movements at the frame 3006 (circled with red) caused the error in the estimation of the acceleration of the feature points at the frame 3007. Thereby, the errors have larger effect on the second-order prediction.

Tai chi chuan is one of the chinese martial arts and the feature is in the slow movement, therefore, the movement and the acceleration of the feature points can easily be detected. Figure 5 shows the tracking result for Tai chi chuan demonstration. The blue pixels represents the predicted object shape (ground truth), the green ones represent the translated predicted shape to determine the object position using (6), and the red ones represents the reconstructed object shape, as calculated by (1). The result image is synthesized from these three shapes, therefore, the white pixels indicate agreement of the both result to the ground truth, the magenta pixels indicate the agreement of the second-order prediction to the ground truth, the cyan pixels indicate the agreement of the linear prediction to the ground truth, and the yellow pixels indicate the agreement of the second-order prediction to the linear prediction but contrary to the ground truth.

Figure 6 shows the chamfer distance to the ground truth, calculated over frames 2950–3050. The result shows that the

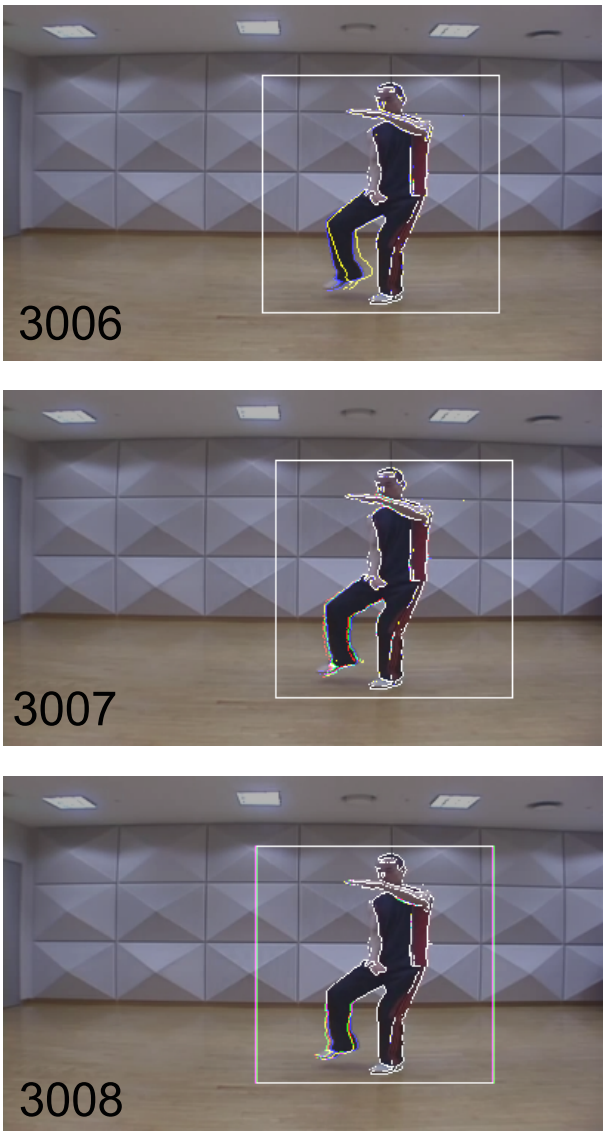


Figure 7. Erroneous video frames for 2nd-order shape prediction of Tai chi chuan.

second-order shape prediction attained better accuracy than the linear one in most of the video frames except 3001 and 3008.

2) Skier 1: backshot: In the video sequence of Tai chi chuan, the object translation is small compared to the object deformation. However, in the sequence of skier captured by a fixed camera, the object translation is much larger than the object deformation. Therefore, the translation and the acceleration of the object might affect the shape prediction accuracy.

Figure 8 shows the tracking result and Figure 9 shows the shape accuracy for the video sequence skier 1. The result in Figure 8, the estimation error (the minimum chamfer distance) tends be high when the object changed its moving direction, such as the video frames around 400, 435, and 475. It also shows that the second-order prediction attained better shape prediction accuracy than linear prediction in most of the video frames, though the second-order prediction produced larger error against the linear method at video frames 478, 479, and

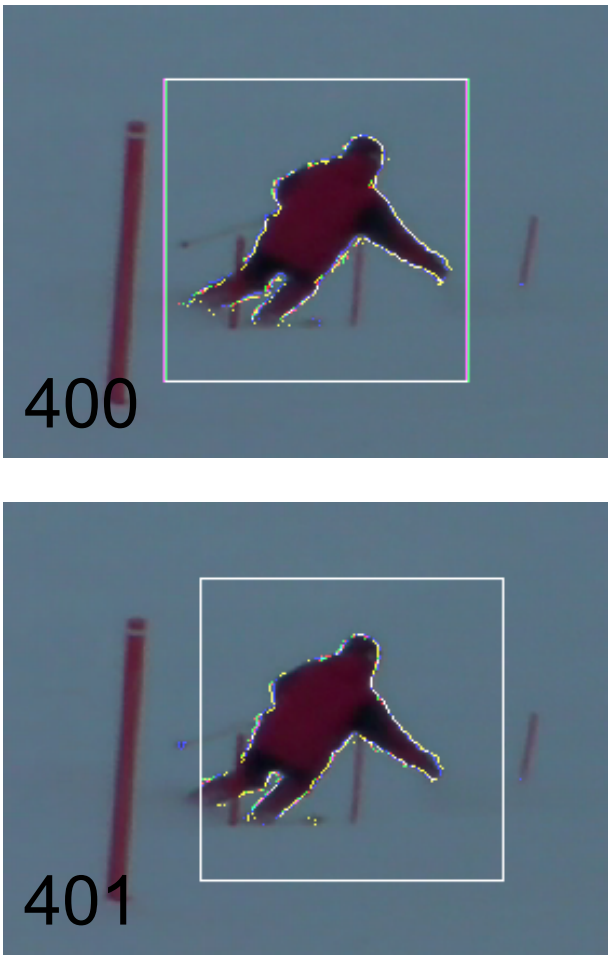


Figure 8. Tracking result for Skier 1 (Fixed camera).
Blue: Ground Truth; Green: Linear Prediction; Red: Second-order Prediction.

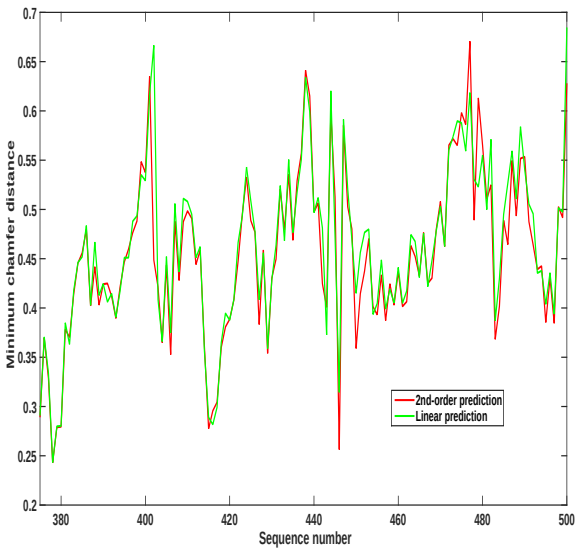
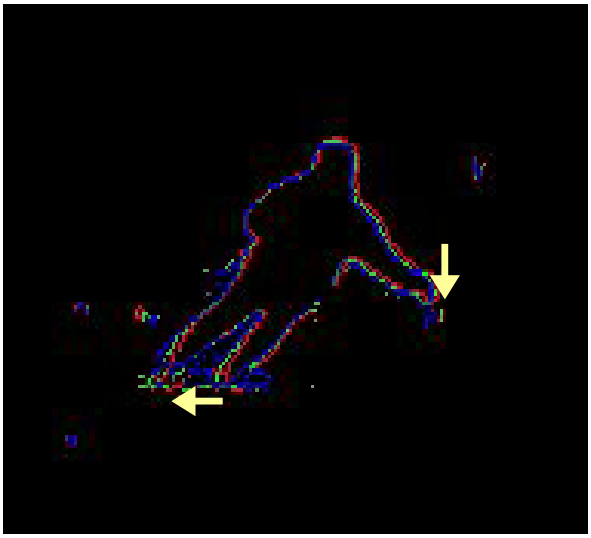


Figure 9. Shape accuracy for Skier 1.

(a) Prediction result for Frame 480



(b) Object movement during frame 478-480

Figure 10. Erroneous video frames for 2nd-order shape prediction of Skier 1 backshot.
(a) Red: 2nd-order prediction; Green: linear prediction; Blue: ground truth.
(b) Red: frame 478; Green: frame 479; Blue: frame 480; Yellow arrow: local movement

480. During video frames 478–480, the object translation was very small and the local movements were also small (Figure 10), therefore the estimation error in local movement (optical flow) must have affected the accuracy of the second-order prediction.

3) Skier 2: frontshot: Figure 11 shows the tracking result and Figure 12 prediction accuracy for skier 2 frontshot. The shape accuracy (Figure 12) shows the same tendency as Figure 8. The estimation error tends be high when the object changed its moving direction, such as the video frames around 240 and 280. In this video sequence, only at the three frames 240, 262 and 288, our second-order method could not outperform the linear method. We considered that the un-eliminated background might affect the prediction accuracy (Figure 13).

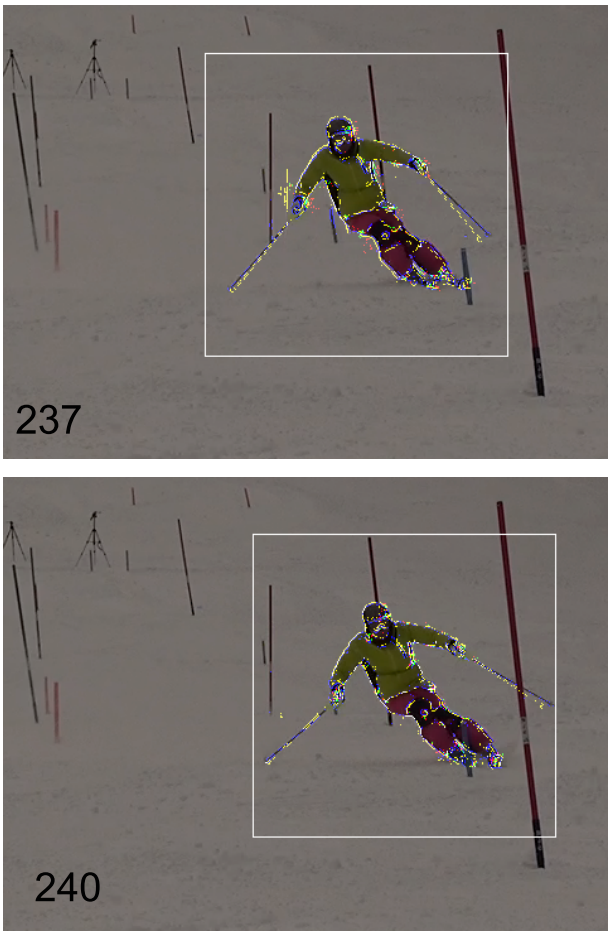


Figure 11. Tracking result for Skier 2 (Fixed camera).
Blue: Ground Truth; Green: Linear Prediction; Red: Second-order Prediction.

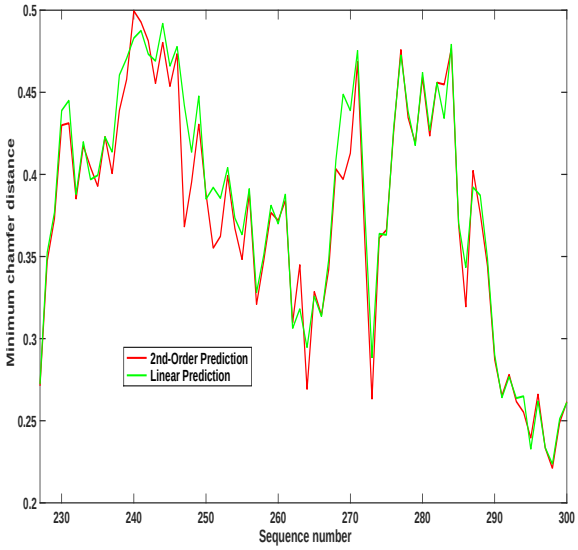


Figure 12. Shape accuracy for Skier 2 by Fixed camera.

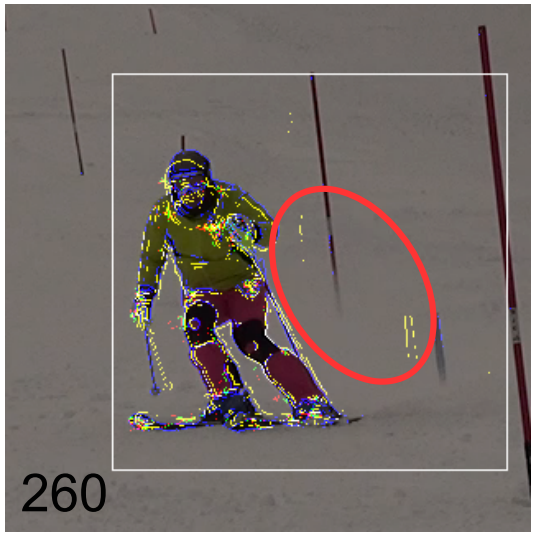
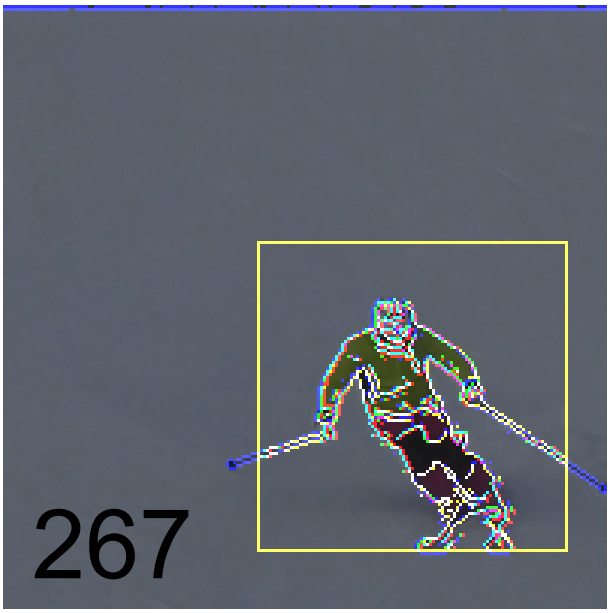


Figure 13. Frame with low shape accuracy for Skier 2.
Un-eliminated background feature points (circles with red) affected the shape and tracking accuracy.



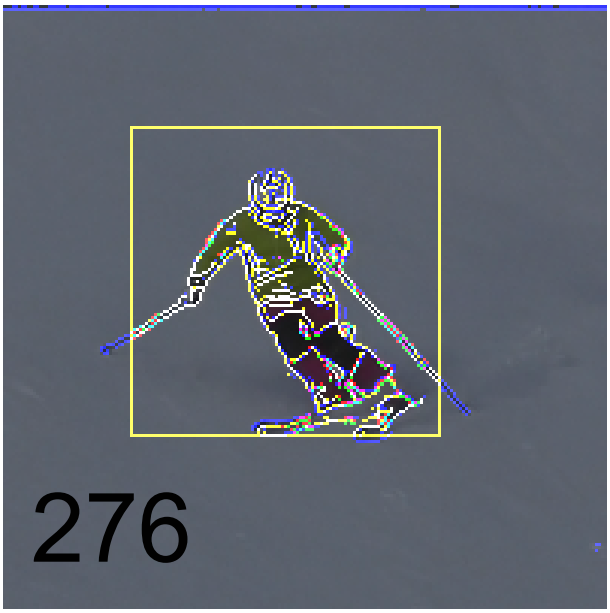


Figure 14. Tracking result for Skier 2 (Hand-held camera.
Blue: Ground Truth; Green: Linear Prediction; Red: Second-order Prediction.

B. Video Sequence captured by Hand-Held Camera (Skier 2)

In the skiing sequence captured by a hand-held camera, the skier was manually "tracked" so as to be shown close to the center of the image frame. Thus, the object tends to exhibit only a small translation in the image frame. However, the object sometimes suffers from a large degree of translation due to manual mis-tracking of the camera. Figure 14 shows the tracking results.

Figure 15 shows the chamfer distance to the ground truth, calculated over frames 230–300. The results show that the second-order prediction attained better accuracy than the linear prediction in 40 out of 70 frames. The second-order prediction is superior during frames 244–249, whereas the linear prediction is preferable from frames 238–240.

Figure 16(a) shows the object translation from frames 244–

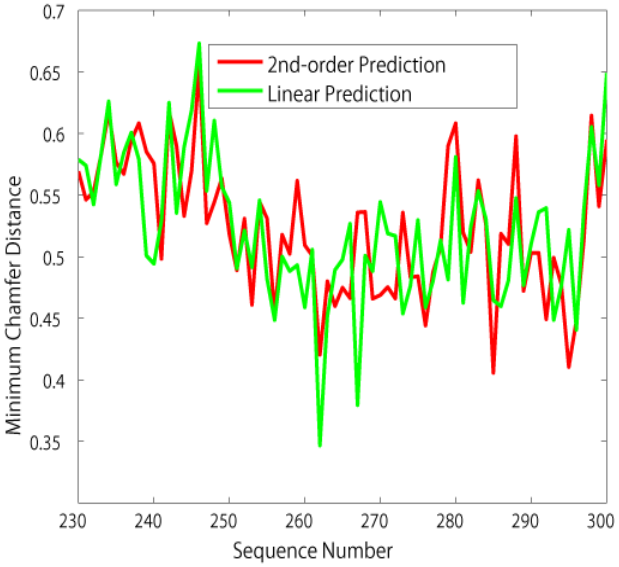


Figure 15. Shape accuracy for Skier 2 by hand-held camera.

248, indicating the direction change at around frame 246. Figure 16(b) shows the object translation from frames 238–240, when the translation direction did not change.

These results indicate that the second-order shape prediction method works well when the direction in which the object must be translated changes.

V. Conclusions

We have evaluated the performance of a second-order shape prediction algorithm. Though the performance is generally higher to that of a linear model, our method outperformed the linear approach in most cases especially when the direction of object movement changed. However, our approach could not outperform the linear approach when the acceleration of the feature points are too high against the frame rate of the video to capture. This evaluation result indicates that the proposed second-order model is robust to objects under acceleration with adequate frame rate.

Acknowledgment

The authors would like to thank Dr. Akaho, group leader of Mathematical Neuroinformatics Group, for his valuable comments and suggestions. This work was supported by JSPS KAKENHI Grant Number 26330217.

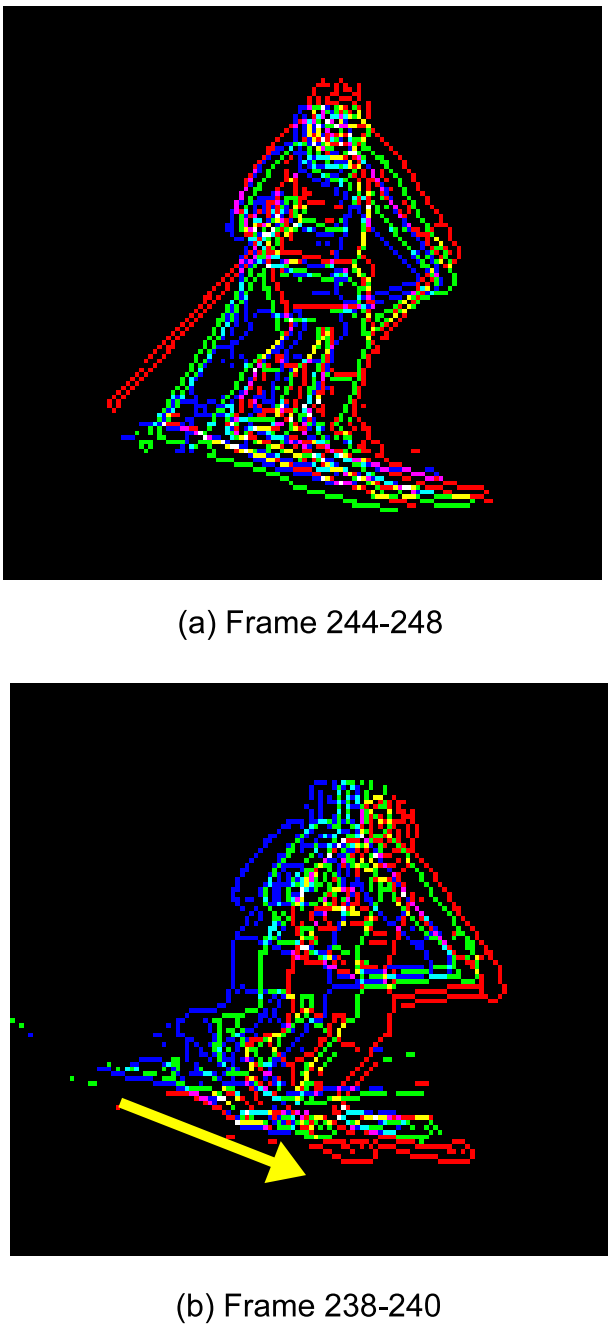


Figure 16. The effect of object translation for prediction accuracy
(a) Blue: frame 244; Green: frame 246; Red: frame 248; Yellow arrow: object translation.
(b) Blue: frame 238; Green: frame 239; Red: frame 240; Yellow arrow: object translation.

REFERENCES

[1] K. Nishida, T. Kobayashi, and J. Fujiki, "The Effect of 2nd-Order Shape Prediction on Tracking Non-Rigid Objects," in *Proc. of the 7th international Conference on Pervasive patterns and Applications*, pp. 60-63, 2015.

[2] G. Sundaramoorthi, A. Mennucci, S. Soatto, and A. Yezzi, "A New Geometric Metric in the Space of Curves, and Applications to Tracking Deforming Objects by Prediction and Filtering," in *SIAM J. of Imaging Science*, Vol. 4, No. 1, pp. 109-145, 2010.

[3] M. Godec, P. M. Roth, and H. Bischof, "Hough-based Tracking on Non-rigid Objects," in *J. of Computer Vision and Image Understanding*, Vol. 117, No. 10, pp. 1245-1256, 2013.

[4] K. Hara, "Real-time Inference of 3D Human Poses by Assembling Local Patches," in *Proc. of IEEE Winter Vision Meeting 2009*, pp. 137-144, 2009.

[5] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient matching of pictorial structures," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2000*, Vol. 2, pp. 66-73, 2000.

[6] K. F. Sim and K. Sundaraj, "Human Motion Tracking of Athlete Using Optical Flow & Artificial Markers," in *Proc. of International Comference on Intelligent and Advanced Systems (ICIAS) 2010*, pp. 1-4, 2010.

[7] K. Nishida, T. Kobayashi, and J. Fujiki, "Tracking by Shape with Deforming Prediction for Non-Rigid Objects," in *Proc. of International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pp. 581-587, 2014.

[8] D. Koller, J. Weber, and J.Malik, "Robust Multiple Car Tracking with Occlusion Reasoning," in *Proc. of European Conference on Computer Vision (ECCV)*, Vol. A, pp. 189-196, 1994.

[9] D. Beymer, P. McLauchlan, B. Coifman, and J. Malik, "A Real-Time Computer Vision System for Measuring Traffic Parameters," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 1997*, pp. 495-501, 1997.

[10] B. Coifman, D. Beymer, P. McLauchlan, and J. Malik, "A Real-time Computer Vision System for Vehicle Tracking and Traffic surveillance," in *Transportation Research Part C: Emerging Technologies*, Vol. 6, No. 4, pp. 271-288, 1998.

[11] Z. W. Kim and J. Malik, "Fast Vehicle Detection with Probabilistic Feature Grouping and its Application ot Vehicle Tracking," in *Proc. of 9th International Conference on Computer Vision (ICCV)*, pp. 524-531 2003.

[12] D. Comaniciu and P. Meer, "MeanShift: A Robust Approach Toward Feature Space Analysis," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 5, pp. 603-619, May, 2002.

[13] D. Comaniciu, V. Ramesh, and P. Meer, "Real-Time Tracking of Non-Rigid Objects using Mean Shift," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2000*, pp. 142-149, 2000.

[14] S. Avidan, "Ensemble Tracking," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 2, pp. 261-271, 2007.

[15] H. Grabner, M. Grabner, and H. Bischof, "Real-Time Tracking via On-line Boosting," in *Proc. of British Machine Vision Conference (BMVC)*, pp. 47-56, 2006.

[16] R.T. Collins, Y. Liu, and M. Leordeanu, "Online Selection of Discriminative Tracking Features," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 10, pp. 1631-1643, 2005.

[17] V. Mahadevan and N. Vasconcelos, "Salliency-based Discriminant Tracking," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognitio (CVPR) 2009*, pp. 1007-1013, 2009.

[18] H. Grabner, C. Leistner, and H. Bischof, "Semi-Supervised On-Line Boosting for Robust Tracking," in *Proc. European Conference on Computer Vision (ECCV) 2008*, pp. 234-247, 2008.

[19] T. Woodley, B. Stenger, and R. Chipolla, "Tracking using Online Feature Selection and a Local Generative Model," in *Proc. of British Machine Vision Conference (BMVC) 2007*, pp. 86.1-86.10, 2007.

[20] D. M. Gavrila, "Pedestrian Detection from a Moving Vehicle," in Proc. European Conference on Computer Vision (ECCV), 2009, pp. 37-49.

[21] D. Huttenlocher, G. Klanderman, and W. J. Rucklidge, "Comparing Images using the Hausdorff Distance," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 9, pp. 850-863, 1993.