Pedestrian Action Prediction using Static Image Feature

Kenji Nishida[1], Takumi Kobayashi[1], Taro Iwamoto[2] and Shinya Yamasaki[2]
[1]*National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba Ibaraki, 305-8568, Japan*
[2]*Mazda Motor Co., 2-5 Moriya-cho Kanagawa-ku, Yokohama, Kanagawa 221-0022, Japan*

Keywords: Action Prediction, Feature Selection, Intelligent Transport System, Image Feature Extraction.

Abstract: In this study, we propose a method to predict how the pedestrian move (run or walk) in the future by using only appearance-based image features. Such kind of motion prediction significantly contributes to intelligent braking system in cars; knowing that the objects will run in several seconds such as for crossing streets, the car can start braking in advance, to effectively reduce the risk for crash accidents. In the proposed method, we empirically evaluate which frames preceding the target action, 'running' in this case, are effective for predicting it in the framework of feature selection. By using the most effective frames at which the image features are extracted, we can build the action prediction method. In the experiments, those frames are found around 0.37 second before running action and we also show that they are closely related to human motion phases from walking to running via biomechanical analysis.

1 INTRODUCTION

According to Japanese traffic accident statistics (Ishikawa, 2010), the number of pedestrian accidents are not decreasing while total number of accidents are decreasing. Moreover, the fatality rate in the pedestrian accidents are five times higher than the other accidents. Therefore, prevention of the pedestrian accidents is one of the most urgent issue in our society. The statistics (Ishikawa, 2010) also shows that 70% of the fatal pedestrian accidents occurred during crossing streets, and thus it is particularly important to safely detect/recognize those crossing pedestrians.

The fatality risk of pedestrian accidents is actually affected by the impact speed (Rosen and Sander, 2009): it is about 4% at the impact speed of 40km/h while it increases to about 10% at 50km/h and 20% at 60km/h. Thus, roughly speaking, the fatal risk decreases by 10% as the impact speed decreases by 10km/h. In the situation that automatic emergency braking (AEB) system works on $6m/s^2$ as defined by Euro-NCAP (Hulshof et al., 2013), it also means that if a car brakes 0.5 sec earlier, the fatality risk in pedestrian accidents would be decreased by 10%. For realizing early braking, it is not sufficient only to detect pedestrians, but it is highly requited to recognize the pedestrian action of high risk, such as crossing street with running, as early as possible.

In the last decade, pedestrian detection is one of the most successful applications in the computer vision and pattern recognition fields. For example, Dalal and Triggs attained over 99% detection rate by introducing HOG feature (Daral, 2005), and very recently it is further improved by deep CNN (Ouyang et al., 2015). However, as described above, just detecting pedestrians is not sufficient for reducing the risk of pedestrian-car accidents. Keller and Gavrila detected crossing people by analyzing pedestrian movement which can be distinguished by the trajectory in the feature space (Keller and Gavrila, 2014). Although they showed promising results such as the accuracy of 0.8 in classifying the correct pedestrian action about 570 ms before the event, it is generally difficult to estimate the precise movement of pedestrians from on-board camera due to its self-motion (shaking). Reddy and Krishnaiah focused on running pose to detect the pedestrian crossing streets (Reddy and Krishnaiah, 2013). These approaches detect the change of pedestrian action from walking to running with the accuracy of 92%, but the detection is performed after the pedestrian already starts running, which is considered to be too late to contribute toward early braking.

We tackle a challenging problem to predict a high risk human action before it actually occurs. In the

Nishida, K., Kobayashi, T., Iwamoto, T. and Yamasaki, S..
Pedestrian Action Prediction using Static Image Feature.
In *Proceedings of the 7th International Joint Conference on Computational Intelligence (IJCCI 2015) - Volume 3: NCTA*, pages 99-105
ISBN: 978-989-758-157-1

realistic situations, we have to pay careful attention to the pedestrians that cross a street with suddenly running and such (sudden) running is regarded as a high risk action to be treated by AEB system with early braking. Therefore, in this paper, we address the problem to predict the (sudden) running action of pedestrians by detecting the *sign* for that action which *preindicates* the running actions beforehand. In addition, we employ an appearance-based approach using only *static image features*, though motion features might be suitable for recognizing actions, since it is quite hard to extract reliable motion features from a moving on-board camera. There is a primary question how early we can predict the running action, or more basically, whether such sign (*preindicator*) exists or not. We empirically answer this question in the framework of feature selection and show the effective preindicator from the quantitative viewpoint. Besides, we also give useful qualitative meaning to it from the biomechanical viewpoints.

2 APPEARANCE BASED ACTION PREDICTION

In this section, we detail the action prediction method using only *static image features*. This method is based on the assumption that the action preindicator can be sufficiently described by distinctive pedestrian shape, not motion itself.

2.1 Static Image Feature

To characterize the human shape in detail, we employ gradient local auto-correlation (GLAC) method (Kobayashi and Otsu, 2008). The GLAC method extracts co-occurrence of gradient orientation as second-order statistics while HOG (Daral, 2005) is based only on first-order statistics of occurrence of gradient orientations. Suppose the pedestrian is detected by arbitrary methods and the bounding box enclosing the pedestrian is provided as shown in fig. 1. As in the common approach such as of HOG (Daral, 2005), the bounding box is spatially partitioned into regular grids of 3×3 at each of which the GLAC features are extracted, then the final feature vector is constructed by concatenating those feature vectors; the setting of 9 orientation bins for gradients and 4 spatial co-occurrence patterns produces GLAC features of 324 dimensionality, and the final feature is formed as a $2916 = 324 \times 3 \times 3$ dimensional vector.

The spatial grids of 3×3 is much coarser compared to HOG-related methods. The GLAC method can characterize the human shape more discriminatively due to exploiting co-occurrence and thus even such coarser grids are enough for static image features. In addition, the coarser grids render robustness regarding spatial position of human shape; that is, the features are stably extracted even for miss-aligned bounding boxes. On the other hand, 3×3 grids are considered as the coarsest one for capturing the human shape; head, torso, two arms and two legs are roughly aligned to respective spatial grids.

2.2 Action Prediction

Based on the time-series sequence of image features extracted in the bounding boxes, we predict the action which will occur in the near future.

We consider the subsequence of T frames which are represented by image feature vectors as described in the previous subsection. Then, we pick up D frames (feature vectors) from them, $[t-D+1, t]$, to predict the action which will occur at the T-th frame indexed as time 0. Those D feature vectors are concatenated to single feature vector of relatively high dimension (fig. 2) which is finally passed to a linear SVM classifier for predicting whether running will occur at time 0 or not. The concatenated feature indirectly encodes motion information of pedestrian during D frames. Because we can not know which timing $\{t, D\}$ produces better performance for predicting the running action, those parameters are empirically determined based on data from the quantitative viewpoint. And, it is obvious that the smaller t is preferable since it provides the earlier prediction; on the contrary, $t = 0$ means on time classification and does not give any prediction at all.

3 EXPERIMENTS

This section shows the experimental procedure for determining the parameters $\{t, D\}$ in the proposed method (section 2.2) as well as evaluating it.

3.1 Dataset

The dataset that we use contains 57 video sequences of 12 children captured by a (fixed) video camera with 30 fps in a gymnasium (fig. 3). [1] Children behave *unpredictably* in context and thus are regarded as the subjects to be carefully paid attention. They first walk

[1] This experiment is approved by the Ethical Review Board of Mazda Motor Corporation and the informed consent of all subjects were also obtained.

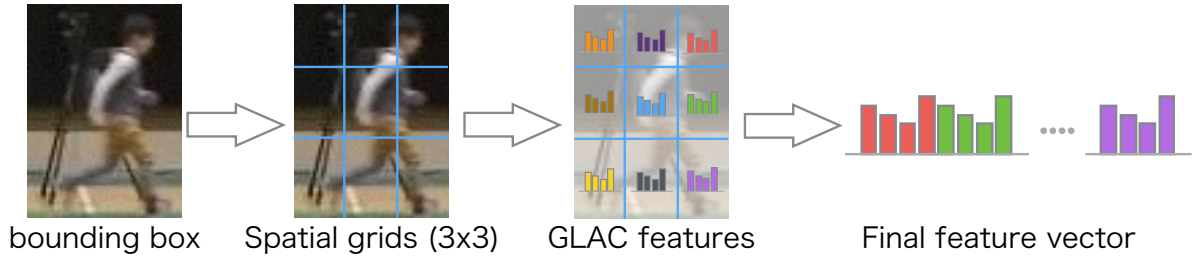


Figure 1: Static image feature extraction by using GLAC method (Kobayashi and Otsu, 2008). The bounding box is partitioned into 3×3 regular grids at each of which GLAC image feature is extracted.

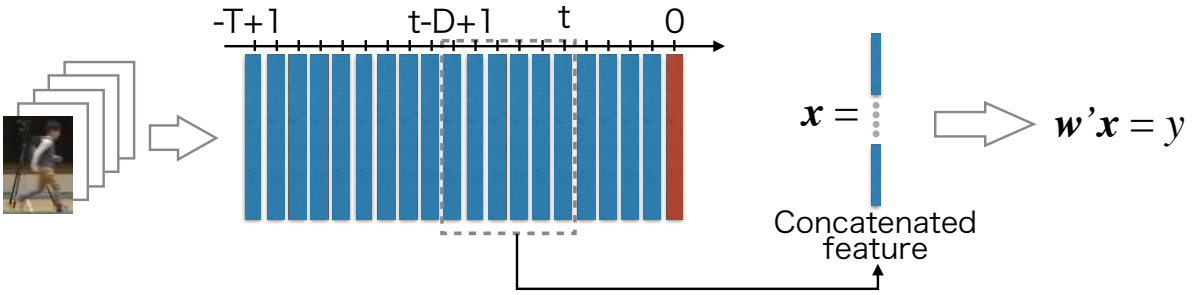


Figure 2: Action prediction framework. We consider T-frame subsequence as a unit. The action of running at $t = 0$ is predicted by using D-frame features preceding it.

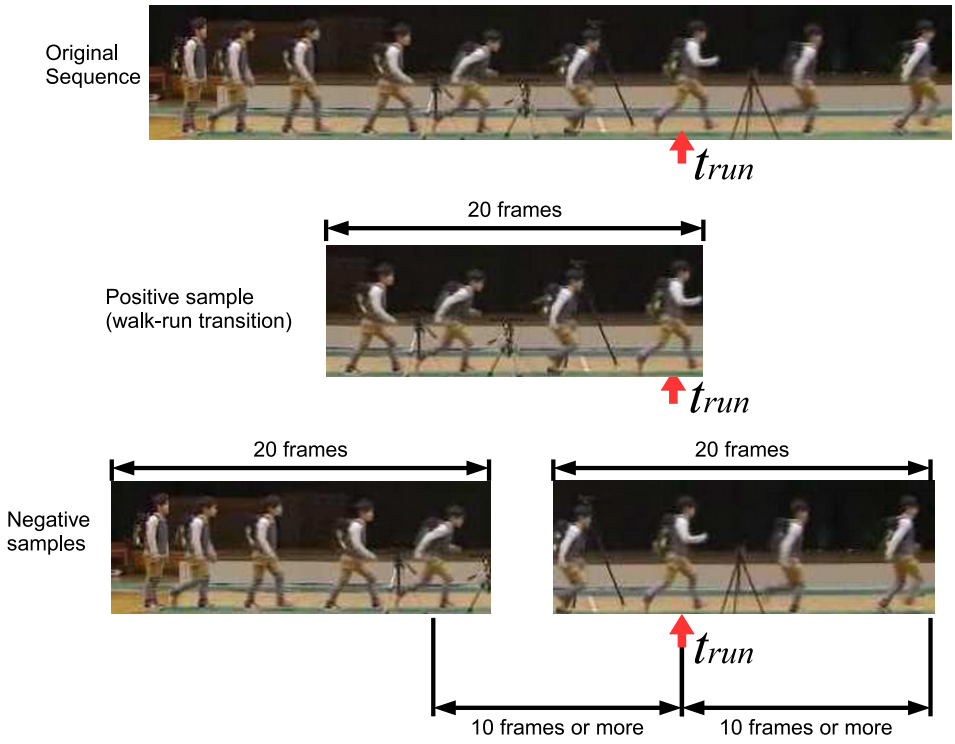


Figure 3: Example sequence in dataset and extracted samples.

and then suddenly run in an arbitrary timing. The bounding boxes enclosing them are manually annotated since the pedestrian detection is out of our focus in this study. In addition, the frame when the subject starts running is also manually indicated; it is denoted as t_{run} (fig. 3). The length of the subsequence is set to $T = 20$, since all the subjects of 57 sequences are *definitely* walking at the frame of $t_{run} - 19$; so the sign

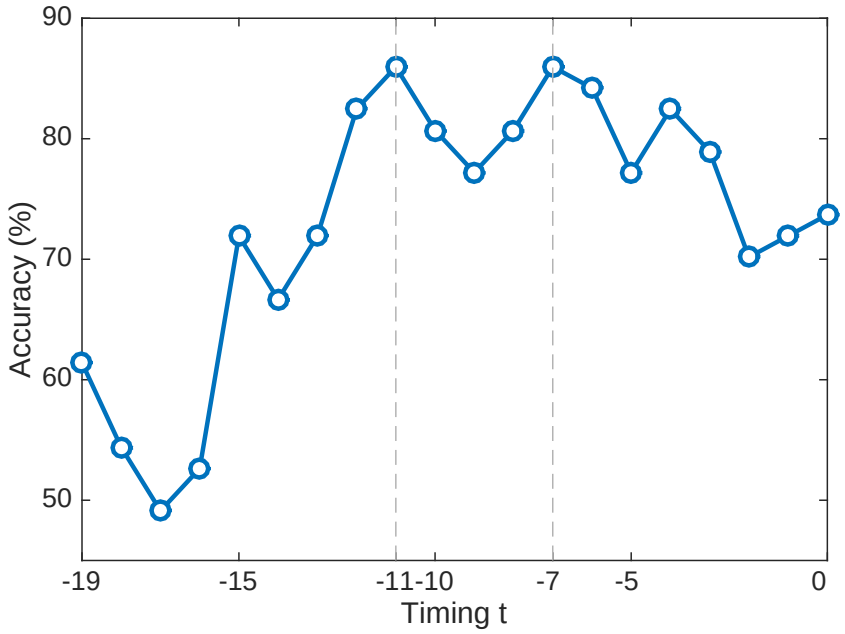


Figure 4: Classification performance of one frame duration $D = 1$.

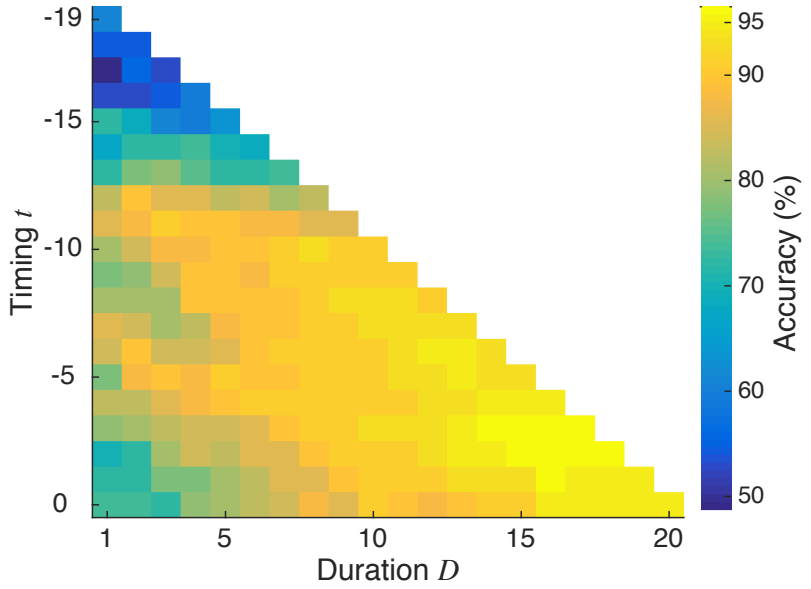


Figure 5: Classification performance for all parameter pairs.

preindicating running is supposed to exist within this period from $t_{run} - 19$ to t_{run}.

The subsequence of 20 frames that ends at t_{run} is regarded as a *positive* sample in each subsequence, while *negative* samples are determined as all the other subsequences except the ones overlapping the positive subsequence with over 10 frames. We thereby obtain one positive sample and about 50 to 100 negative samples from each sequence.

3.2 Evaluation

The prediction performance is measured by *leave-one-sequence-out cross validation*, the procedure is defined as follows. At the i-th iteration ($i = 1, .., 57$), we train the linear SVM classifier (Vapnik, 1998) over the samples excluding the ones drawn from the i-th sequence. Then, the samples from the i-th sequence are evaluated by applying the classifier. In this case, those evaluated samples are highly imbalanced due to

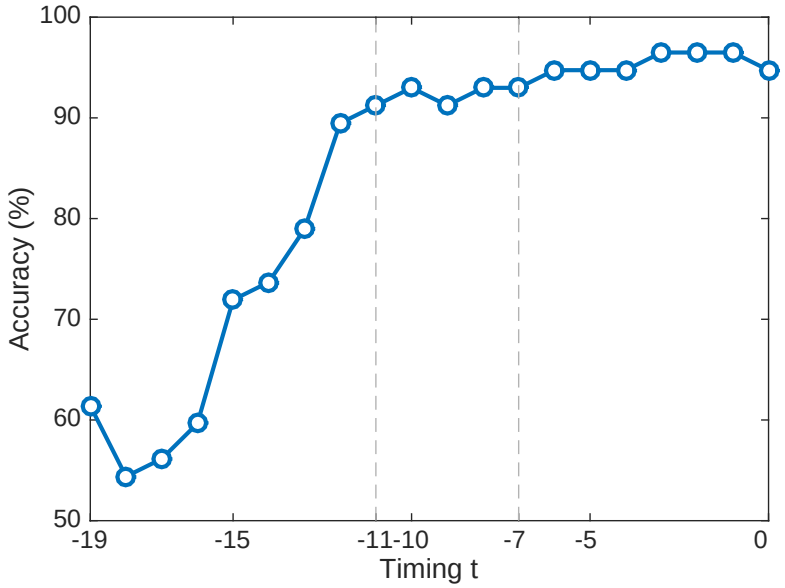


Figure 6: Classification performance for timing t with maximizing over D.

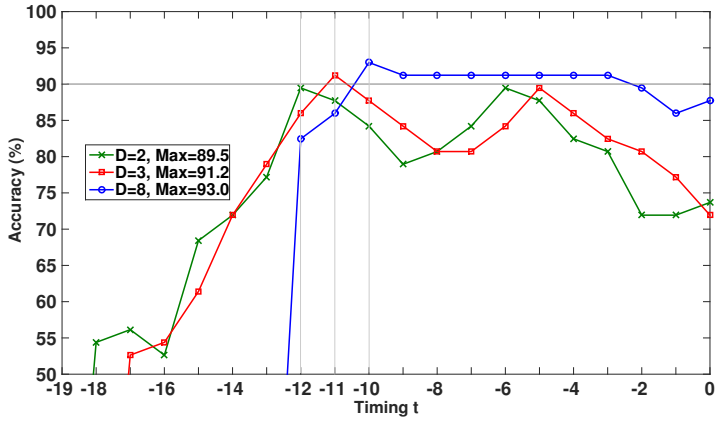


Figure 7: Classification performance for duration $D = 2,3,8$.

containing only one positive sample. Therefore, we regard the i-th sequence as correctly classified only when all the sample of that sequence are successfully classified, which is a relatively hard criterion. In an overall evaluation, we measure the ratio of the correctly classified sequences out of 57 ones.

As to the prediction method (sec.2.2), we examined 210 pairs of $\{t,D\}$ parameters: considering $T = 20$, the prediction timing t varies from 0 to -19, and accordingly the period D can be changed from 1 to $t+20$.

4 EXPERIMENTAL RESULTS

Figure 4 shows the classification performance for one frame duration ($D = 1$). We can see that the top accuracy was obtained at $t = -11$ and -7. This result suggests that the frames at $t = -11$ and -7 include distinctive features to preindicate running. It should be noted that though this task is to predict the running action at $t = 0$, the performance at $t = 0$ (on-time classification) is not high. This is because the pedestrians definitely run at $t = 0$ and some negative samples also contain the running action at $t = 0$, making hard to classify at $t = 0$.

Figure 5 shows the results for all parameter pairs of $\{t,D\}$. The best accuracy 96.5% was attained at $t = -3$ with $D = 14,16$ and $t = -2$ with $D = 18$; the whole sequence ($t = 0, D = 20$) did not perform the best, exhibiting 94.7%. However, $t = -2$ and $t = -3$ are not preferable for our purpose, early prediction.

As shown in fig. 4, the distinctive features are found at $t = -7$ and -11, and thus we can push back

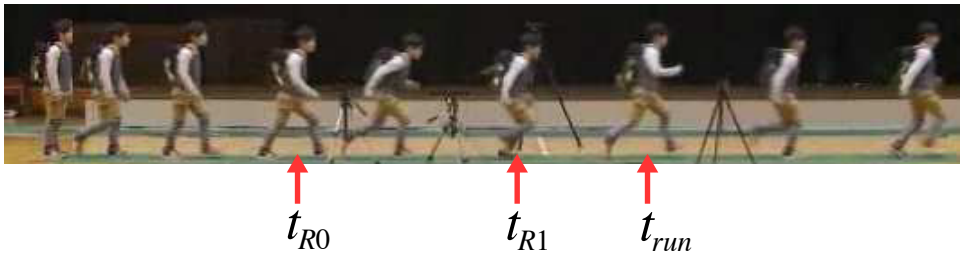
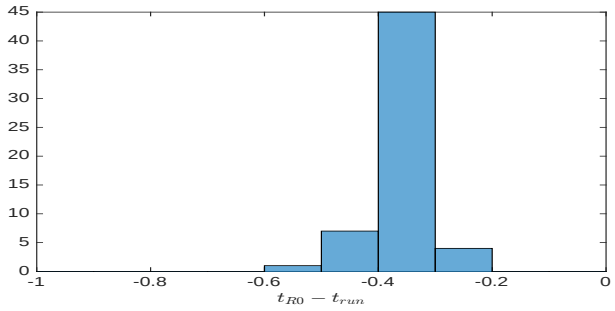
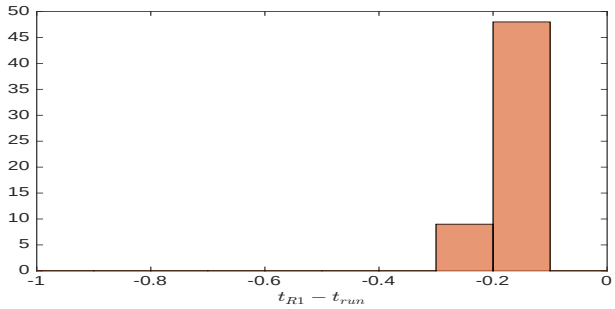


Figure 8: Biomechanical analysis for transition from walking to running.



(a) t_{R0}



(b) t_{R1}

Figure 9: Histogram for t_{R0} and t_{R1} compared to t_{run}

the prediction earlier. For early prediction, the timing t is rather important than the duration D, and we show in fig. 6 the best performance at each t by picking up the maximum accuracy over D. It apparently shows that the performance is saturated at $t = -11$, slightly increasing after $t = -11$; for example, 93.0% is attained at $t = -10$ with $D = 8$ which is close to the best 96.5% at $t = -3$. However, a pedestrian has to be tracked through 8 frames for the duration $D = 8$, which is not preferable for on-board (moving) cameras. Figure 7 shows the classification accuracy and timing with the duration $D = 2, 3$ and 8. 91.2% is attained at $t = -11$ with $D = 3$ which requires only 3 frame duration. Thus, we can conclude that it is possible to predict the action of running at about 0.37 sec. earlier (corresponding to $t = -11$) with over 90% accuracy. Moreover, if we can compromise with the classification accuracy of 89.5%, the running action can be predicted at about 0.4 sec. earlier ($t = -12$) by using only 2 frame duration.

5 BIOMECHANICAL ANALYSIS

During the transition from walking to running, the visually most distinctive form is found when the head reaches the minimum height. After that, the pedestrian jumps up a little bit and subsequently the phase is completely changed into running. We call this point as t_{R1} (see fig. 8). On the other hand, when a pedestrian starts running from walking, the form accordingly changes in order to facilitate its acceleration. At that point, the pedestrian's posture is learning forward as well as stepping and shaking the arms more largely. This point is denoted as t_{R0} (see fig. 8). The form at t_{R0} is less salient compared to that at t_{R1}, but t_{R0} precedes t_{R1}. For comparing the above results (section 4) to these biomechanically distinct points,

we manually annotated t_{R0} and t_{R1} in the sequences. The histograms for those timing points are shown in fig 9. Those timing points are not diverse across the pedestrians but relatively concentrated around the means. This result shows that those distinct points defined from the biomechanical viewpoint are also regarded as general measure for predicting running action. Those means are $\bar{t}_{R0} = t_{run} - 0.37$ sec. and $\bar{t}_{R1} = t_{run} - 0.19$ sec., corresponding to $t = -11$ and $t = -6$, respectively. These are surprisingly coincident with the points $t = -11$ and -7 which are qualitatively obtained in fig. 4. Thus, we have shown that those quantitatively obtained timing points are biomechanically meaningful.

6 CONCLUSION

We have proposed a method to predict the running action of pedestrians at earlier timing before the action actually occur. The method is based on the appearance-based image features to extract distinctive forms of the pedestrian in transition from walking to running. In addition, the motion information is naively encoded via aggregating (concatenating) the consecutive frame features in time series sequence, with the two important parameters which indicate the timing and duration, respectively. In the experiments, we empirically determined those two parameters, showing favorable performance of prediction; the running action can be predicted at about 0.4 sec before. By further analyzing the postures from the viewpoint of biomechanics, the prediction timing is shown to be closely related to biomechanically distinct form. The experiments performed in this paper are limited due to such as indoor and fixed camera. Our future works include to apply the proposed method to the movie which is captured in more realistic situations.

REFERENCES

T.Ishikawa: The analysis of pedestrian accidents. https://www.itarda.or.jp/ws/pdf/h22/13_01 hokousyaziko.pdf in japanese

Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 886–893 (2005)

Hulshof, W., Knight, I., Edwards, A., Avery, M., Grover, C.: Autonomous emergency braking test results. In: International Technical Conference on the Enhanced Safety of Vehicles (2013)

Keller, C.G., Gavrila, D.M.: Will the pedestrian cross? a study on pedestrian path prediction. IEEE Transaction on Intelligent Transportation Systems 15(2), 494–506 (2014)

Kobayashi, T., Otsu, N.: Image feature extraction using gradient local auto-correlations. In: European Conference on Computer Vision. pp. 346–358 (2008)

Ouyang, W., Wang, X., Zeng, X., Qiu, S., Luo, P., Tian, Y., Li, H., Yang, S., Wang, Z., Loy, C.C., Tang, X.: Deepid-net: Deformable deep convolutional neural networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2403–2412 (2015)

Reddy, Y., Krishnaiah, R.: Driving assistance system for identification of sudden pedestrian crossings. International Journal of Research in Information Technology 1(12), 281–295 (2013)

Rosen, E., Sander, U.: Pedestrian fatality risk as a function of car impact speed. Accident Analysis and Prevention 41, 536–542 (2009)

Vapnik, V.: Statistical Learning Theory. Wiley (1998)