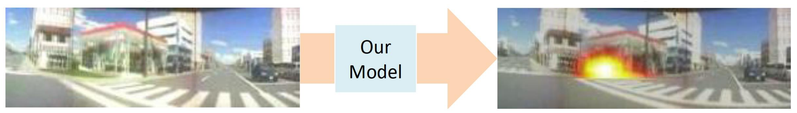
画像変換に基づく高齢者の運転時注視マップの推定

オンカー クリシュナ、入江 豪、川西 隆仁、柏野 邦夫 (NTT)、相澤 清晴 (東大)

Abstract

Task: Predicting which part of a scene elderly people would pay attention, especially in driving scenario.

Driving scene → Our Model → Elderly's FoA



Importance: As of in 2018, Japan had 5.63 million drivers aged 75. Can we assist them to avoid the fatal accidents?

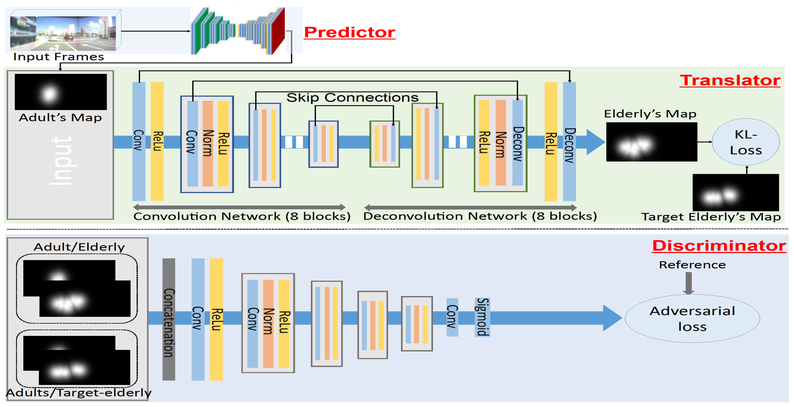
Proposal: First deep image-to-image translation method for predicting elderly focus of attentions (FoAs)

Model

Existing models: can precisely predict adult's FoA, however, they do not work well for elderly's.

Idea: can we leverage the knowledge of the FoA predictors constructed for the adults to predict the elderly's?

Architecture: our model mainly consist of two major networks; translator (T) and discriminator (D).



Translator has simple encoder-decoder architecture, whereas discriminator is consist of encoder.

Training

Overall objective is to train the two networks that solve the following optimization problem

$$\min_D \max_T \mathcal{L}_{\mathrm{Adv}}(T, D) - \gamma \mathcal{L}_{\mathrm{KL}}(T)$$

It consist of two major loss terms; content loss $\mathcal{L}_{\mathrm{KL}}$ and adversarial loss $\mathcal{L}_{\mathrm{Adv}}$

$$\mathcal{L}_{\mathrm{KL}}(T) = \sum_n \sum_i e^*_{n,i}\left(\log\left(e^*_{n,i}\right) - \log\left(e_{n,i}\right)\right)$$

$\mathcal{L}_{\mathrm{KL}}$ requires T to output the ground truth FoA $e^*_{n,i}$ in per pixel bases, $\mathcal{L}_{\mathrm{Adv}}$ is to facilitate the training of the translator.

Experiment Setup

We assume a car driving scenario for our experiment, the setup consist of driving simulator with eye-gaze trackers.

Experiment setup (Driving simulator)

18 observers belonging to adults (mean age 26) and elderly (75 years) age groups participated in the study.

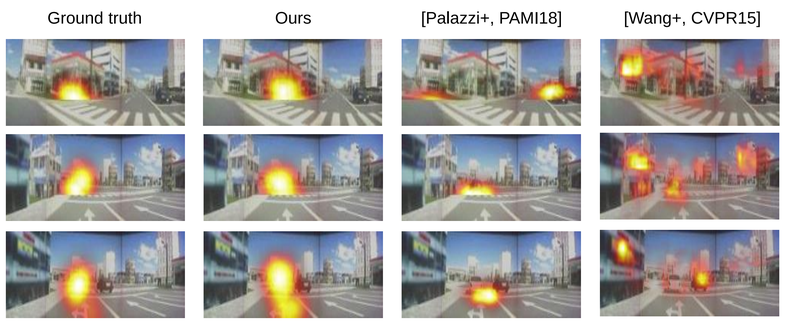
Adult and eldery's FoA obtained is shown in the above figure. 7,722 images used for training and 1,930 for testing

Results

Quantitative result is measured by correlation coefficient (CC), similarity measure (SIM), and KL-divergence

Algorithm	CC ↑	SIM ↑	KL-div.↓
Ours	**0.7717**	**0.6832**	**2.18**
[Palazzi+, PAMI18]	0.6386	0.5324	4.06
[Palazzi+, PAMI18] (fine-tuned)	0.6575	0.5535	-
[Wang+, CVPR15]	0.1305	0.2232	5.60
[Wang+, TIP15]	0.0901	0.2595	4.90
[Cornia+, ICPR16]	0.1478	0.2940	-

Qualitative result shows that only our model can precisely recover the ground-truth maps of elder's.



Conclusions

We attempted to accurately predict the elderly's FoA by introducing a deep image translation framework.

References

1. Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, "Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model," IEEE Transactions on Image Processing, vol. 27, no. 10, pp. 5142–5154, 2018.

ACKNOLEDGEMENT: We would like to thank Prof. Kimihiko Nakano of The University of Tokyo.