ICIP2021

TIME-LAG AWARE MULTI-MODAL VARIATIONAL AUTOENCODER USING BASEBALL VIDEOS AND TWEETS FOR PREDICTION OF IMPORTANT SCENES

【PAPER ID: 2311】

Kaito Hirasawa, Keisuke Maeda, Takahiro Ogawa and Miki Haseyama *Hokkaido University, Japan*

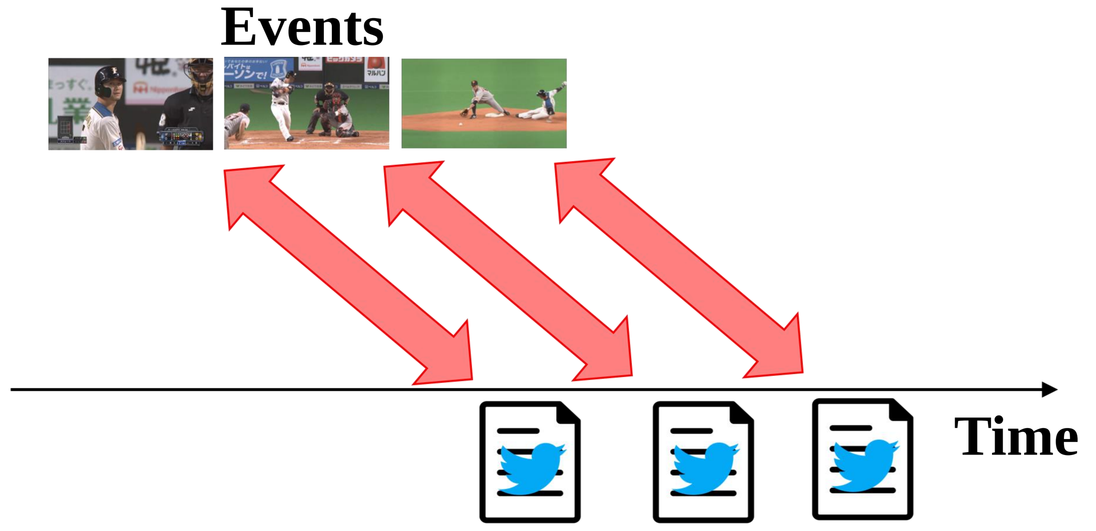
INTRODUCTION

- If viewers know when important scenes of sports videos will occur by the prediction of these scenes, they can efficiently view these scenes at the timing.
- For this prediction, the video-based studies [5-6] and the study utilizing both e-sports videos and audience chat reactions [7] have been researched.

- Tweets posted on Twitter[*1] often include the reactions of the viewers and explain the details of the games.

The construction of a highly accurate method for the prediction of important scenes is expected by using tweets and videos.

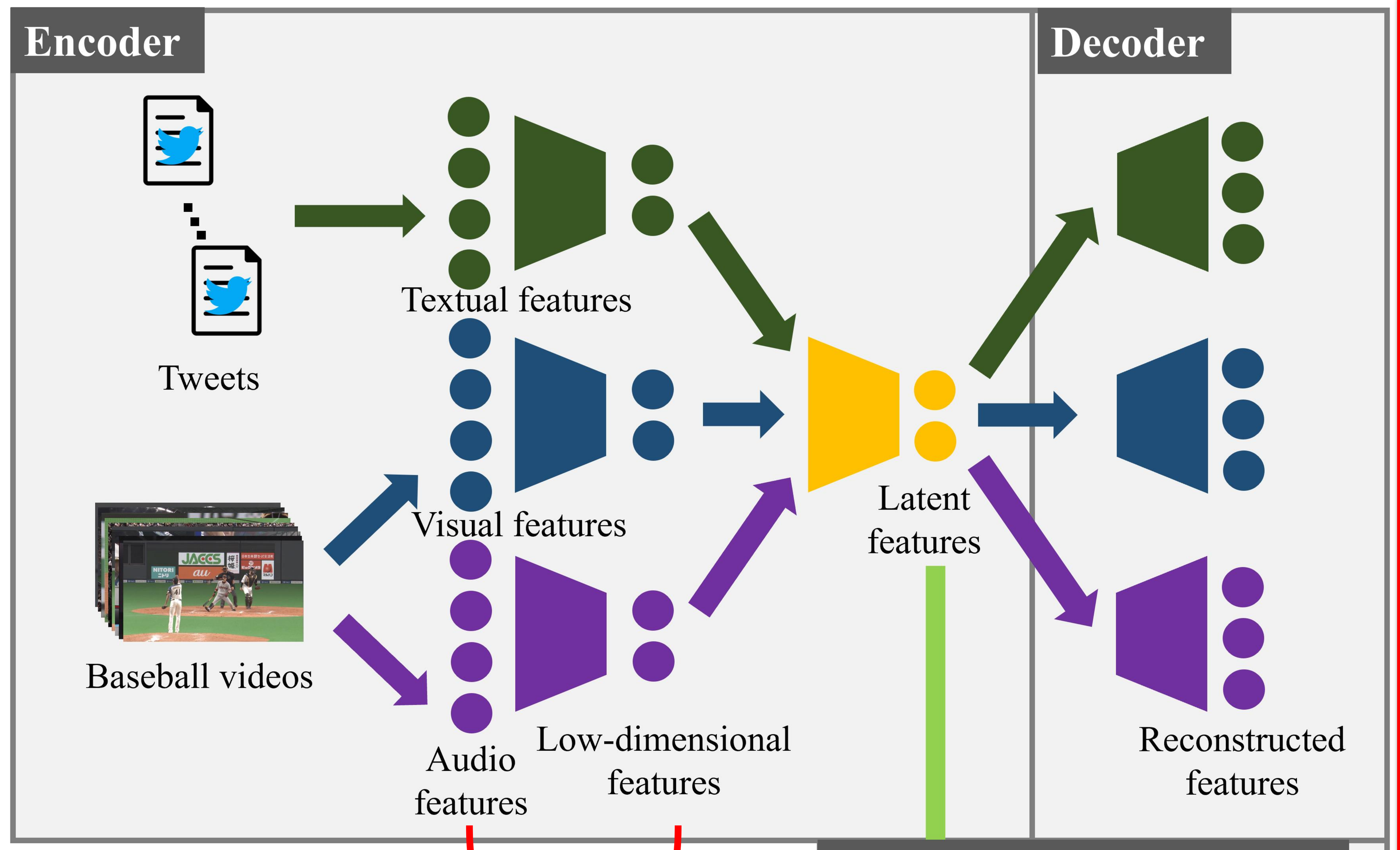
- Since multiple previous events in the videos influence tweets posted on Twitter, they are closely related to each other.
- Thus, there are time-lags between tweets and corresponding multiple previous events.



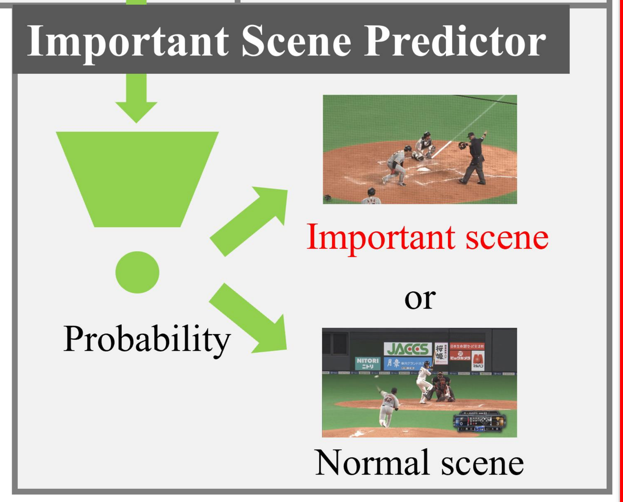
Problem : There are not any methods considering these time-lags for the prediction of important scenes in sports videos.

PROPOSED METHOD

A method via a **time-lag aware multi-modal variational autoencoder [12] for prediction of important scenes (TlMVAE-PIS)** in baseball videos



Contribution



Loss function L_{final}

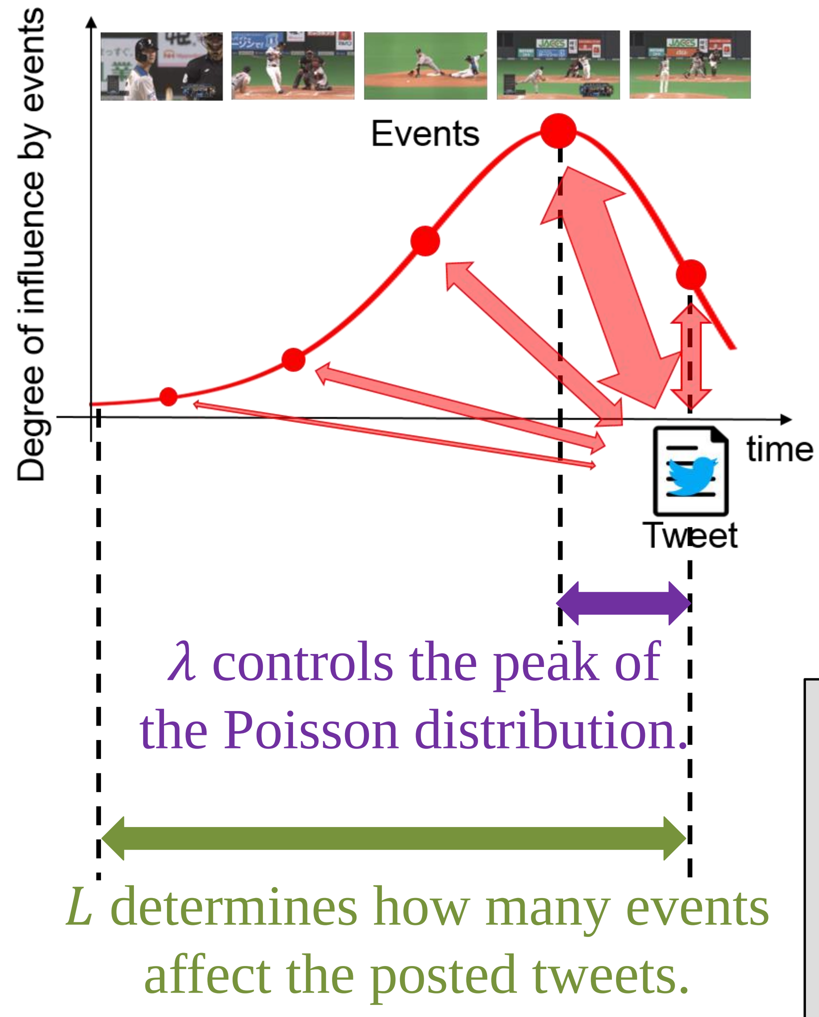
$$L_{\mathrm{final}} = \left(\sum_{m\in\{\mathrm{t,v,a}\}} \xi_{\mathrm{rec}}^{m} L_{\mathrm{rec}}^{m}\right) + \xi_{\mathrm{kl}} L_{\mathrm{kl}} + \xi_{\mathrm{isp}} L_{\mathrm{isp}}$$

$\xi_{\mathrm{rec}}^{m}, \xi_{\mathrm{kl}}, \xi_{\mathrm{isp}}$: Weighting parameters
L_{rec}^{m}: Reconstruction loss
L_{kl} : KL divergence
L_{isp}: Prediction loss of important scenes

- By learning to bring the output of the decoder and the input of the encoder closer together, we can extract the important information needed for the reconstruction.

Contribution

- The influence of the just previous event is strong, and the influence of the past event tends to be gradually weakened.
- Our method assumes that tweets are affected according to the Poisson distribution.



Covariance matrix considering the time-lags R^{m_1,m_2}

$$\boldsymbol{R}^{m_1,m_2} = \begin{cases} \dfrac{\sum_{l=0}^{L-1} \frac{e^{-\lambda}\lambda^{l}}{l!} \widehat{\boldsymbol{X}}_0^{m_1} \widehat{\boldsymbol{X}}_l^{m_2\top}}{\sum_{l=0}^{L-1} \frac{e^{-\lambda}\lambda^{l}}{l!}} & (m_1 \in \{\mathrm{t}\}, m_2 \in \{\mathrm{v,a}\}) \\ \dfrac{\sum_{l=0}^{L-1} \frac{e^{-\lambda}\lambda^{l}}{l!} \widehat{\boldsymbol{X}}_l^{m_1} \widehat{\boldsymbol{X}}_0^{m_2\top}}{\sum_{l=0}^{L-1} \frac{e^{-\lambda}\lambda^{l}}{l!}} & (m_1 \in \{\mathrm{v,a}\}, m_2 \in \{\mathrm{t}\}) \\ \widehat{\boldsymbol{X}}_0^{m_1} \widehat{\boldsymbol{X}}_0^{m_2\top} & (\text{otherwise}) \end{cases}$$

$\widehat{\boldsymbol{X}}_l^{m} = \left[\boldsymbol{x}_{L-l}^{m}, \dots, \boldsymbol{x}_{|W|-l}^{m}\right]$ $(l = 0, \dots, L-1)$
$m, m_1, m_2 \in \{\mathrm{t,v,a}\}$: Modality
t, v, a: Textual, visual, audio
λ: Parameter of the Poisson distribution
L: Parameter determining how many previous events affect the posted tweets
$|W|$: Number of the tweets
$\boldsymbol{x}_i^{m}$: i-th features of modality m

Contribution : Consideration for the time-lags for the derivation of the covariance matrices

EXPERIMENTAL RESULTS

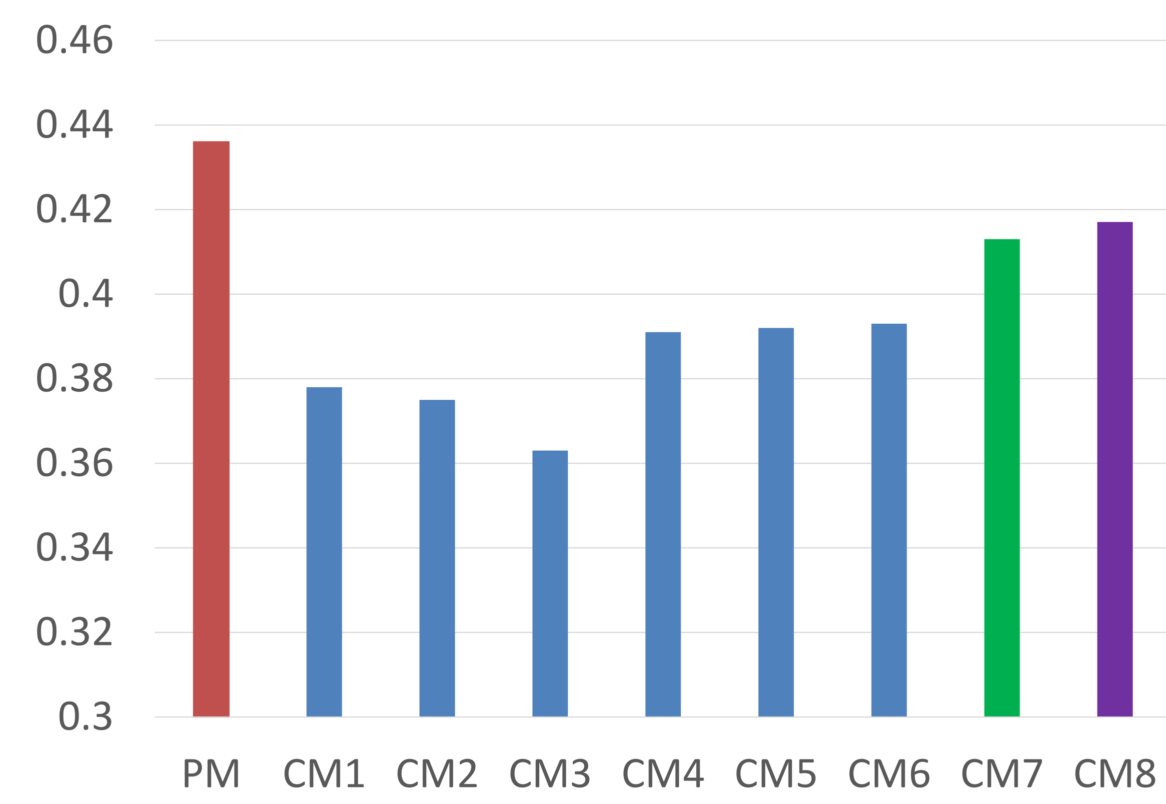
Settings

Dataset : 12 games from June 14th to September 27th in 2019
Seven games as training data and the other five games as test data
Tweets during the games including an official hashtag of the team
Ground truth : The labels given by eight subjects who were healthy males aged between 20 and 24 years with 11-15 years of baseball experience
Evaluation index : F-measure

Comparative methods (CMs)

CMs1-6 : Using features of {textual}, {visual}, {audio}, {textual, visual}, {textual, audio} and {visual, audio}, respectively. CMs4 and 5 consider the time-lags.
CM7 : MVAE [22] not considering time-lags
CM8 : Long Short-Term Memory [25]

Average F-measure in the proposed method (PM) and CMs1-8



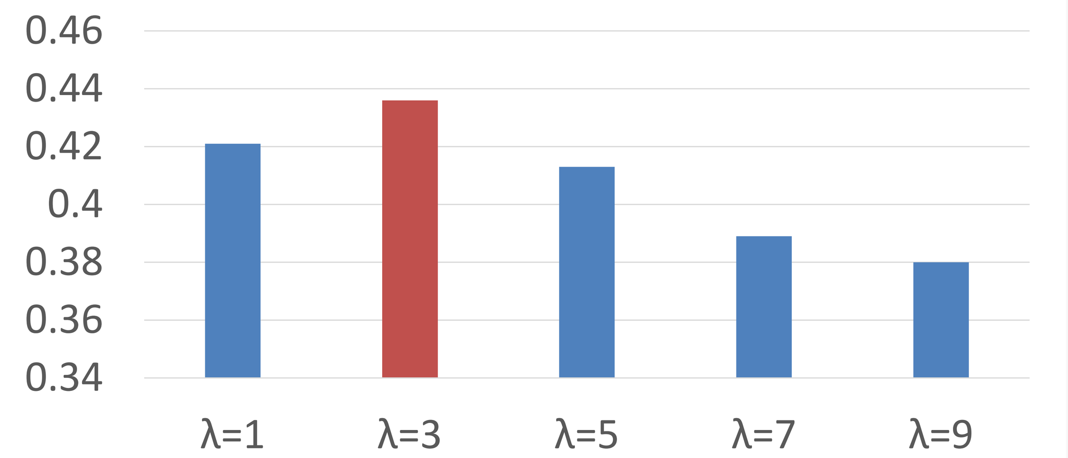
PM vs CMs1-6
- Confirmed the effectiveness of using textual, visual and audio features.

PM vs CM7
- Confirmed the effectiveness of the consideration of the time-lags.

PM vs CM8
- Confirmed the effectiveness of adopting MVAE for the prediction of important scenes.

We verify PM is effective for the important scene prediction.

Average F-measure in PM when changing the parameter λ



- It is confirmed that the highest F-measure is achieved when λ is three.
- The tweet of the test data is posted every 24 seconds on average.

The time-lag between the tweets and the corresponding event is about 72 seconds.

Acknowledgement : This work was partly supported by JSPS KAKENHI Grand Number JP17H01744.

*1 https://twitter.com/