ABSTRAK

Perkembangan dalam integrasi bidang *Computer Vision* dan *Natural Language Processing* telah mendorong kemajuan sistem cerdas berbasis multimodal. Salah satunya penerapannya adalah sistem penerjemah bahasa isyarat. Namun, penelitian pada dataset Bahasa Isyarat Indonesia (BISINDO) masih terbatas, terutama dalam pengembangan sistem *Sign Language Translation* (SLT) yang mampu menerjemahkan video menjadi teks bahasa alami. Keterbatasan dataset, kompleksitas gerakan multimodal, serta minimnya *pipeline* yang mendukung proses penerjemahan menjadi tantangan utama dalam pengembangan sistem tersebut. Oleh karena itu, penelitian ini mengusulkan sebuah *Pipeline Vision-and-Language* untuk menerjemahkan video BISINDO menjadi teks bahasa alami secara otomatis. Penelitian ini mengadaptasi kerangka kerja *Cross-Industry Standard Process for Data Mining* (CRISP-DM) yang mencakup tahapan *business understanding*, *data understanding*, *data preparation*, *modelling*, dan *evaluation*. Pipeline yang dikembangkan terdiri dari ekstraksi fitur berbasis MediaPipe untuk memperoleh *landmark* tangan dan tubuh, pemodelan spasio-temporal menggunakan *encoder* berbasis *Convolutional Neural Network* (CNN) dan *Gated Recurrent Unit* (GRU), serta GRU-*Attention* pada *decoder* untuk menghasilkan representasi teks yang lebih kontekstual. Penelitian ini menggunakan dataset yang terdiri dari 30 kalimat BISINDO dengan total 900 video. Hasil evaluasi menunjukkan bahwa model yang diusulkan mampu mencapai nilai *Bilingual Evaluation Understudy* (BLEU) *score* sebesar 0,42 dan *Word Error Rate* (WER) sebesar 0,31, yang menunjukkan kemampuan model dalam menghasilkan terjemahan yang cukup akurat pada dataset terbatas. Penelitian ini memberikan kontribusi sebagai *baseline* dalam pengembangan sistem penerjemah BISINDO berbasis *deep learning*.

Kata Kunci: Bahasa Isyarat Indonesia, *Deep Learning*, Mekanisme *Attention*, Penerjemahan Bahasa Isyarat, *Vision-and-Language*

ABSTRACT

The advancement in the integration of Computer Vision and Natural Language Processing has driven the development of multimodal intelligent systems. One of its applications is sign language translation systems. However, research on Indonesian Sign Language (BISINDO) datasets remains limited, particularly in developing Sign Language Translation (SLT) systems capable of translating video into natural language text. Challenges such as limited dataset availability, complex multimodal gestures, and the lack of supporting pipelines pose significant obstacles in developing such systems. Therefore, this study proposes a Vision-and-Language Pipeline to automatically translate BISINDO videos into natural language text. This research adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, which includes the stages of business understanding, data understanding, data preparation, modeling, and evaluation. The proposed pipeline consists of feature extraction using MediaPipe to obtain hand and body landmarks, followed by spatio-temporal modeling using an encoder based on Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU), and a GRU-Attention decoder to generate more contextual text representations. The dataset used in this study consists of 30 BISINDO sentences with a total of 900 video samples. The evaluation results show that the proposed model achieves a Bilingual Evaluation Understudy (BLEU) score of 0.42 and a Word Error Rate (WER) of 0.31, indicating that the model is capable of producing reasonably accurate translations on a limited dataset. This study contributes as a baseline for the development of BISINDO translation systems based on deep learning.

Keywords: *Attention Mechanism, Deep Learning, Indonesian Sign Language, Sign Language Translation, Vision-and-Language.*