

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Bahasa isyarat merupakan bahasa yang telah digunakan oleh sebagian besar masyarakat tuna rungu dan tuna wicara di seluruh dunia. Bahasa ini menggabungkan variasi pergerakan tubuh dan tangan sekaligus raut muka sebagai medium komunikasi. Ada banyak variasi bahasa isyarat pada setiap negara, masing-masing mempunyai pergerakan badan dan isyarat visual yang berbeda-beda. Di Indonesia, bahasa isyarat indonesia disebut BISINDO (Bahasa Isyarat Indonesia).

Walaupun bahasa isyarat memungkinkan komunikasi bagi masyarakat disabilitas, masih terdapat permisahan antara pengguna bahasa isyarat dan masyarakat awam yang tidak mengerti bahasa isyarat. Selain itu, jumlah orang yang mengerti dan mengajar bahasa isyarat sangat sedikit dibanding dengan masyarakat awam [1], sehingga masyarakat disabilitas cenderung akan merasa terputus dengan komunikasi dengan dunia luar. Oleh karena itu, sistem prediksi bahasa isyarat otomatis telah dikembangkan untuk mengatasi masalah dalam komunikasi dan edukasi bagi masyarakat awam. Berdasarkan [2], sistem prediksi bahasa isyarat telah menjadi topik riset selama lebih dari 30 tahun.

Belakangan ini, penelitian-penelitian terkait sistem prediksi bahasa isyarat [3] [1] [4] berfokus ke penggabungan beberapa faktor di video bahasa isyarat daripada menggunakan satu tipe jaringan saraf untuk mendapatkan akurasi yang lebih baik dalam memprediksi bahasa isyarat. Penelitian-penelitian ini biasanya menggabungkan fitur-fitur global seperti informasi kerangka tubuh bagian atas (torso) yang didapat menggunakan mesin estimasi pose, orang yang memeragakan bahasa isyarat (instruktur), dan/atau aliran optik yang didapatkan dari algoritma aliran optik seperti TV-L1 atau farneback dengan fitur lokal seperti wajah atau tangan instruktur. Fitur-fitur ini kemudian dilatih secara terpisah, kemudian prediksinya akan dirata-ratakan untuk menghasilkan prediksi akhir. Teknik ini mendapatkan akurasi yang lebih tinggi dibandingkan dengan penelitian sebelumnya yang hanya menggunakan satu fitur global sebagai input.

Penelitian yang memakai teknik tersebut mendapatkan tingkat akurasi yang beragam. Penelitian maruyama et al [1] menggabungkan 6 fitur model yaitu fitur global berupa video instruktur dan aliran optik instruktur, fitur lokal berupa tangan

dan muka, dan fitur kerangka tubuh dengan dataset Word-Level American Sign Language (WLASL) [5] dan MS-ASL [6] yang masing-masing mempunyai 2000 kata dan 1000 kata dalam bahasa isyarat Amerika dan mempunyai akurasi top-1 sebesar 47.26% pada dataset WLASL dan 49.06% pada dataset MS-ASL. Penelitian Özdemir et al [3] menggabungkan 4 fitur model yaitu fitur muka, kerangka tubuh, dan video tangan dengan dataset BosphorusSign22k [7] dan AUTSL [8], masing-masing mempunyai 744 dan 226 kata dalam bahasa isyarat Turki dan mempunyai akurasi top-1 sebesar 92.58% pada dataset BosphorusSign22k dan 90.85% pada dataset AUTSL. Penelitian Shin et al [4] menggabungkan fitur kerangka tubuh dan kerangka tangan dengan dataset *Korean Sign Language* (KSL) yang terdiri dari 77 kata bahasa isyarat Korea dan mempunyai akurasi sebesar 99.87%.

Sedangkan penelitian sistem prediksi BISINDO belakangan ini tampaknya hanya berfokus ke prediksi kategori alfabet dan angka dari dataset gambar. Meskipun penelitian ini mempunyai akurasi yang signifikan dalam memprediksi bahasa isyarat alfabet dan angka, penelitian tersebut hanya menyediakan subset yang kecil dibandingkan kosakata bahasa isyarat Indonesia. Telah terdapat pula prediksi bahasa isyarat Indonesia dalam level kata. Penelitian yang dilakukan aljabar et al [9], memprediksi 8 kata dan 2 alfabet BISINDO yang mempunyai akurasi sebesar 90%. Penelitian yang dilakukan handhika et al [10] memprediksi 25 kata BISINDO dan memperoleh akurasi sekitar 60%. Walaupun penelitian ini menghasilkan akurasi yang baik, penelitian ini hanya menggunakan dataset dengan jumlah kata yang sedikit (≤ 25). Oleh karena itu, dibutuhkan penelitian lebih lanjut untuk mengembangkan sistem prediksi bahasa isyarat menggunakan dataset BISINDO yang lebih luas untuk mendapatkan jangkauan kata yang lebih banyak.

Terinspirasi dari penelitian maruyama et al [1], yang menggabungkan data global (aliran optik, video instruktur bahasa isyarat), data lokal (tangan dan muka), dan data pose (berupa titik-titik pose tubuh) maka penelitian ini akan menerapkan metode yang telah dilakukan pada penelitian tersebut pada BISINDO. Namun, penelitian ini tidak memakai model lokal dikarenakan akurasi pada kombinasi model yang menggunakan model lokal dengan akurasi yang tidak menggunakan model lokal tidak sepadan dikarenakan pemakaian model lokal menambah jumlah model dan teknik preproses sebesar 2 kali lipat, sementara akurasi hanya naik sebesar 2%. Oleh karena itu, penelitian ini hanya berfokus ke model video instruktur (rgb), aliran optik, dan pose.

Model	RGB	Flow	Local	Skeleton	MS-ASL100*			MS-ASL200*			MS-ASL500*			MS-ASL1000*		
					Top 1	Top 5	Top 10	Top 1	Top 5	Top 10	Top 1	Top 5	Top 10	Top 1	Top 5	Top 10
Baseline1 [8]	✓				73.69	91.38	93.87	69.16	86.26	90.43	50.27	75.39	82.29	35.10	59.61	68.93
Baseline2		✓			82.46	94.66	96.61	79.02	92.46	95.01	63.36	84.64	88.51	50.59	73.89	80.08
Cross-domain [1]	✓				83.04	93.46	-	80.31	91.82	-	-	-	-	-	-	-
SignBERT(H+P) [60]				✓	81.37	93.66	-	77.34	91.10	-	-	-	-	59.80	80.94	-
SignBERT(H+R) [60]	✓			✓	89.56	97.36	-	86.98	96.39	-	-	-	-	71.24	89.12	-
Ours1	✓		✓		75.12	91.33	93.87	73.11	89.27	92.38	56.37	78.00	83.76	39.55	63.09	70.49
Ours2	✓			✓	75.61	92.38	95.07	70.17	86.27	91.02	52.63	75.81	82.96	35.61	59.50	67.93
Ours3	✓		✓	✓	76.69	92.65	95.35	73.32	89.42	92.18	58.60	79.18	85.28	40.23	63.96	70.92
Ours4	✓	✓	✓		83.84	94.69	96.18	80.40	93.29	95.11	64.66	84.69	88.94	50.59	72.33	78.78
Ours5	✓	✓	✓	✓	84.22	94.77	96.48	78.86	92.26	94.58	63.44	83.97	88.68	48.68	70.84	77.54
Ours6	✓	✓	✓	✓	83.86	94.86	96.66	80.72	92.76	95.43	65.46	84.85	88.94	49.06	70.46	76.97

Gambar 1.1. Perbandingan akurasi antara penelitian yang dilakukan maruyama et al dengan penelitian sebelumnya

Sumber: [1]

Penelitian ini juga akan melakukan beberapa perubahan pada tahap praproses. Menurut [11], akurasi MediaPipe [12] dalam mendeteksi titik pose lebih baik daripada akurasi OpenPose [13][14] yang digunakan pada penelitian Maruyama et al. oleh karena itu, Prediksi pose yang sebelumnya menggunakan OpenPose, diubah ke MediaPipe. Kemudian menurut [15], penggunaan algoritma aliran optik farneback [15] lebih cocok untuk pendeteksi gerakan tubuh walaupun akurasi TV-L1 [16] lebih tinggi dibandingkan Farneback. Oleh karena itu, penelitian ini juga memeriksa penggunaan algoritma farneback untuk kalkulasi aliran optik dibanding dengan algoritma TV-L1 yang digunakan sebelumnya. Praproses deteksi titik-titik video instruktur pada dilakukan dengan menggunakan YOLOv11 [17] dibandingkan penelitian sebelumnya yang menggunakan YOLOv3[18] dikarenakan akurasi dan efisiensi YOLOv11 lebih baik dibandingkan YOLOv3 [17].

1.2 Rumusan Masalah

1. Bagaimana cara untuk memprediksi kata BISINDO dengan menggunakan gabungan model I3D dan ST-GCN?
2. Seberapa akurat prediksi BISINDO dengan menggunakan gabungan dari model prediksi video I3D dan model pose ST-GCN?

1.3 Batasan Permasalahan

1. video dataset terdiri dari satu orang instruktur yang memperagakan sebuah kata BISINDO per video.

2. dikarenakan pada dataset terdapat video duplikat (versi video mirror, kualitas rendah/tinggi, dan ekstensi duplikat), maka video akan difilter dengan mencari versi dengan ukuran penyimpanan terbesar, diasumsikan bahwa video tersebut mempunyai kualitas dan durasi yang paling tinggi.
3. Objek yang diprediksi berupa bahasa isyarat berlevel kata, sehingga tidak memakai dataset huruf.
4. teknik praproses video, *splitting dataset*, dan nilai *epoch* menggunakan penelitian yang dilakukan maruyama et al [1], dikarenakan inspirasi arsitektur penelitian menggunakan metode maruyama et al.

1.4 Tujuan Penelitian

1. Untuk membuat model yang dapat memprediksi BISINDO menggunakan video
2. Untuk memeriksa akurasi pengembangan dari metode penelitian sebelumnya terhadap BISINDO

1.5 Manfaat Penelitian

1. Mempermudah komunikasi antar kaum disabilitas dengan masyarakat awam dengan memperkenalkan sistem pendeteksi kata BISINDO
2. Mengembangkan jangkauan kosakata prediksi BISINDO dari yang sebelumnya kurang dari 25 kata menjadi 50 kata.

1.6 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
Membahas tentang latar belakang permasalahan dan penelitian serta batasan-batasan yang akan ditetapkan pada penelitian ini.
- Bab 2 LANDASAN TEORI
Berisi tentang landasan teori yang akan dijelaskan secara mendetail yang digunakan pada penelitian ini

- Bab 3 METODOLOGI PENELITIAN

Berisi teknik-teknik dan metode yang digunakan pada penelitian.

- Bab 4 HASIL DAN DISKUSI

Berisi hasil dari penelitian dan penjelasan hasil penelitian

- Bab 5 KESIMPULAN DAN SARAN

Berisi kesimpulan dari hasil yang telah dihasilkan selama penelitian sekaligus saran-saran yang dapat digunakan pada penelitian selanjutnya

