INTISARI

Ensemble Model Berbasis Fitur *Spatiotemporal Handcrafted* dan *Deep Visual–Skeleton* untuk Pengenalan Pukulan Badminton

Oleh

Farida Asriani

(22/499014/SPA/00845)

Pengenalan pukulan badminton secara otomatis merupakan tantangan penting dalam bidang *human action recognition* (HAR). Identifikasi pola gerakan, kecepatan permainan, serta perbedaan karakteristik individu antar atlet masih terbatas pada penggunaan satu jenis data visual. Permasalahan lain yang ada adalah tidak meratanya akurasi pada tiap kelas teknik pukulan. Kondisi ini membatasi potensi penerapan praktis dalam dunia nyata seperti analisis performa atlet, sistem pelatihan cerdas, maupun aplikasi sport analytics.

Penelitian ini mengusulkan model yang mengintegrasikan fitur spasial dan temporal dari data video. Metode *ensemble* dikembangkan untuk meningkatkan akurasi model. Tahap awal dimulai dengan preprocessing video menjadi frame representatif. Ekstraksi fitur *handcrafted* dari frame dalam RGB menghasilkan fitur HOG, HOF, dan MBH. Frame-frame RGB juga diolah dengan mediapipe untuk menghasilkan, koordinat skeleton 3D yang menjadi fitur spatial ROMI dan temporal DTW. Proses seleksi fitur dilakukan pada HOG, HOF, MBH, ROMI, dan DTW untuk memperoleh representasi yang lebih optimal. Vision Transformer (ViT) digunakan untuk menghasilkan fitur deep visual. Kombinasi beberapa fitur digabungkan melalui *concatenation* untuk membentuk *hybrid* fitur. Hybrid fitur dengan akurasi terbaik digunakan dalam klasifikasi. Proses klasifikasi dilakukan melalui tiga strategi utama: (1) Ensemble ML yang mengintegrasikan RF, LR, SVM, dan AdaBoost melalui *weighted soft voting* dengan input dari *concatenation* fitur *handcrafted* spatiotemporal, (2) LSTM untuk menangkap dinamika temporal dari fitur hibrid ViT–skeleton, dan (3) Ensemble ML–DL berbasis *weighted soft voting classifier* guna meningkatkan akurasi dan kestabilan prediksi antar kelas pukulan badminton.

Berdasarkan pengujian dari dataset video pukulan badminton, tiga strategi utama memberikan kontribusi signifikan dalam meningkatkan akurasi klasifikasi teknik pukulan badminton Pada strategi pertama model *ensemble* ML mencapai akurasi sebesar 98,21%. Model menggabungkan SVM LR, dan RF dengan data (ROMI, DTW, HOF). Pada strategi kedua model *hybrid* yang mengintegrasikan fitur Vision Transformer (ViT) dan skeleton 3D ke dalam arsitektur LSTM mencapai akurasi sebesar 98,68%. Pada strategi ketiga, diperoleh akurasi tertinggi sebesar 99,69% dari model *ensemble* ML–DL yang menggabungkan LSTM dan RF melalui strategi soft voting classifie. Hasil ini membuktikan bahwa integrasi fitur spasial-temporal

handcrafted dan *deep visual-skeleton* serta kombinasi model ML dan DL mampu meningkatkan performa klasifikasi dan mengatasi permasalahan tidak meratanya akurasi klasifikasi tiap kelas.

Kata kunci : Pukulan badminton, spatiotemporal, handcrafted, deep visual-skeleton, *machine learning*, voting *classifier*, *ensemble* learning.

ABSTRACT

Ensemble Model Based on Spatiotemporal Handcrafted and Deep Visual–Skeleton Features for Badminton Stroke Recognition

By

Farida Asriani

(22/499014/SPA/00845)

Automatic recognition of badminton strokes represents a significant challenge in the field of human action recognition (HAR). The identification of motion patterns, game speed, and individual differences among athletes remains limited by reliance on a single type of visual data. Another critical issue lies in the imbalance classification accuracy across different stroke categories. These limitations restrict the potential for practical applications in real-world contexts, such as athlete performance analysis, intelligent training systems, and sport analytics.

This study proposes a model that integrates spatial and temporal features from video data. An ensemble method is developed to improve model accuracy. The process begins with video preprocessing into representative frames. Handcrafted feature extraction from RGB frames produces HOG, HOF, and MBH features. The RGB frames are also processed using MediaPipe to generate 3D skeleton coordinates, which are then used to derive the spatial feature ROMI and the temporal feature DTW. Feature selection is applied to HOG, HOF, MBH, ROMI, and DTW to obtain a more optimal representation. A ViT is employed to extract deep visual features. Several features are then combined through concatenation to form hybrid features. The hybrid features with the highest accuracy are used for classification. Classification is carried out using three main strategies: (1) an ensemble ML model that integrates RF, LR, SVM, and AdaBoost through weighted soft voting with inputs from concatenated handcrafted spatiotemporal features, (2) an LSTM model to capture temporal dynamics from hybrid ViT–skeleton features, and (3) an ML–DL ensemble using a weighted soft voting classifier to enhance accuracy and prediction stability across badminton stroke classes.

Based on experiments with the badminton stroke video dataset, the three main strategies made significant contributions to improving classification accuracy. In the first strategy, the ensemble ML model achieved an accuracy of 98.21%. This model combined SVM, LR, and RF using ROMI, DTW, and HOF features. In the second strategy, the hybrid model that integrated ViT features and 3D skeleton data into the LSTM architecture reached an accuracy of 98.68%. In the third strategy, the highest accuracy of 99.69% was obtained from the ML–DL ensemble model, which combined LSTM and RF through a soft voting classifier. These results demonstrate that the integration of handcrafted spatiotemporal features and deep visual–skeleton features, along with the combination of ML and DL models, effectively enhances classification performance and addresses the issue of uneven class-wise accuracy.

Keywords: *Badminton stroke technique, spatiotemporal, handcrafted, deep visual-skeleton, machine learning*, voting *classifier*, *ensemble* learning.