Efficient 3Dconv Fusion of RGB and Optical Flow for Dynamic Hand Gesture Recognition and Localization

Gibran Benitez-Garcia[1](✉) and Hiroki Takahashi[1,2]

[1] Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan
gibran@ieee.org, rocky@inf.uec.ac.jp

[2] Artificial Intelligence eXploration Research Center (AIX), Meta-Networking Research Center (MEET), The University of Electro-Communications, Tokyo, Japan

Abstract. Hand Gesture Recognition (HGR) has been significantly advanced through multimodal approaches utilizing RGB and Optical Flow (OF). Yet, two main challenges often remain (i) The computational burden triggered by advanced techniques which rely on intricate multi-level fusion blocks distributed across the architecture, and (ii) the limited exploration into the impact of OF estimators on multimodal fusion. To address these, this paper introduces an efficient RGB+OF fusion relying on just a few 3DConv layers applied early in the architecture. Concurrently, we explore the impact of five state-of-the-art OF methods on this fusion. Advancing beyond traditional HGR, we prioritize recognizing and precisely localizing the hand gesture, which is critical for a wide range of computer vision applications. Thus transitioning the focus to Hand Gesture Recognition and Localization (HGRL). Accordingly, we employ a YOLO-based architecture renowned for its real-time efficacy and precision in object localization, aligning with the demands of dynamic gestures often seen in HGRL. We evaluate our approach with the IPN-Hand dataset, augmenting its scope for HGRL evaluation by manually annotating 82,769 frames. Our experiments show significant results of 10% enhancement in mAP against the RGB-only method and a 7% gain over 2DConv-based fusion.

Keywords: Hand Gesture Recognition and Localization · RGB+Optical Flow Fusion · YOLO-based Architecture

1 Introduction

Automatic Hand Gesture Recognition (HGR) is critical in developing intuitive human-computer interfaces since it focuses on interpreting user hand movements as instructions or commands [1,6]. However, when the crucial aspect of spatial localization is included in the process, we transition to the more comprehensive

Supported by a Research Grant (S) at Tateisi Science and Technology Foundation.

W. Q. Yan et al. (Eds.): PSIVT 2023, LNCS 14403, pp. 197–210, 2024.
https://doi.org/10.1007/978-981-97-0376-0_15

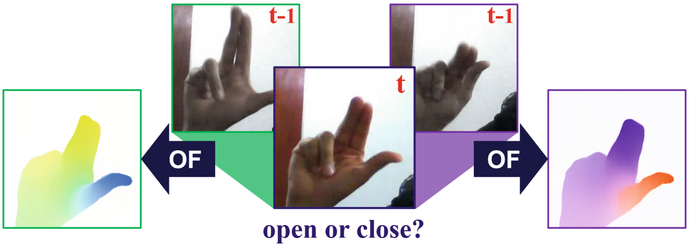

Fig. 1. Illustration of the limitations of relying only on the current frame (t) for classifying dynamic gestures. Temporal information from previous frames must be considered to discern if the fingers are opening (right) or closing (left). Dense Optical Flow (OF) can effectively capture and depict the crucial temporal features for HGRL.

challenge of Hand Gesture Recognition and Localization (HGRL). In HGRL, the gesture and precise hand location play an essential role in a wide range of applications in the automotive sector, virtual reality, industrial electronics, and others [6,23,25]. For instance, touchless screen manipulation is a technology that has become increasingly relevant in a world focused on hygiene and reduced physical contact [1,16]. For such interfaces, while simple commands might be captured through static hand gestures relying solely on spatial cues, interpreting more complex, dynamic gestures need motion interpretation. As depicted in Fig. 1, dynamic gestures are ambiguous when viewed as a single frame, underscoring the importance of temporal cues for HGRL.

Dense Optical Flow (OF) has traditionally been a standard method to extract temporal features for HGR. Several deep learning approaches in the literature have fused the complementary nature of RGB and OF data to create robust multimodal features [7,12,13,17,20]. However, two critical challenges often remain under-addressed in most RGB+OF approaches. Firstly, the computational cost of multimodal architectures frequently goes overlooked. So, the necessity to compute dense OF as a preprocessing step significantly limits their application. Secondly, the accuracy and inference speed performance of OF estimators become crucial to the success of the entire system. Therefore, these challenges must be carefully addressed to design efficient and reliable multimodal HGRL fusion approaches.

The fusion of RGB and OF data is commonly achieved through different techniques, with middle multi-level fusion standing out as it effectively captures low-level and high-level correlations between modalities through the whole Convolutional Neural Network (CNN) [7,12,19]. However, traditional fusion methods rely on fusion blocks to merge modal-specific features, often increasing computational costs. In contrast, in this paper, we introduce an efficient RGB+OF fusion purely based on a couple of 3DConv layers from the early stages of the architecture. This allows the network to holistically learn the complementary multimodal characteristics in an integrated manner rather than generating isolated features that need subsequent fusion. To further optimize our approach, we adopt a YOLO-based single-stage architecture chosen for its real-time processing and exceptional accuracy in spatially localizing objects, aligning perfectly with the demands of dynamic HGRL tasks. This configuration ensures end-to-end learning while maintaining a mix of initial 3D and subsequent 2D convolutions for computational efficiency. On the other hand, the contribution of different OF methods for multimodal HGR has been barely investigated. Therefore, we explore the impact of five state-of-the-art (SOTA) OF methods, specifically in early RGB+OF fusion.

We evaluate our proposed approach using 11 distinct gesture classes of the IPN-Hand dataset [2], known for its challenging dynamic gestures tailored for interactions with touchless devices. While this dataset offers temporal annotation of the gestures, hand location is not provided. Therefore, we manually annotate 82,769 frames, adding another dimension to the dataset, enabling it to evaluate HGRL approaches. Through comprehensive experiments and an in-depth ablation study, we validate the effectiveness of our proposed RGB+OF approach. The results show a notable 10% boost in mAP compared to the RGB-only strategy and a significant 7% increase over a 2Dconv-based fusion. Note that this performance is achieved with a marginal increase in the computational cost of about 1 GFLOP to the baseline architecture. Testing code, pre-trained models, and extended annotations of the IPN-Hand dataset will be publicly available at https://github.com/GibranBenitez/IPN-hand/.

In resume, the main contributions of this paper include:

- Introduction of an efficient 3Dconv fusion of RGB+OF data employing a YOLO-based single-stage architecture for dynamic hand gesture recognition and localization (HGRL).
- Comprehensive analysis and evaluation of five SOTA OF methods: RAFT [22], GMA [10], KPA [15], SKF [21], and FlowFormer [9] for HGRL based on RGB+OF.
- Extension of the IPN-Hand dataset annotations of about 82K frames resulting in 83,613 annotated hands.

2 Related Work

The explosion of deep learning in the last decade urged several solutions for HGR, capitalizing on the advancements of CNNs [1,13,26] and their 3D counterparts, 3D-CNNs [7,12,17]. Preliminary techniques for RGB+OF HGR predominantly adopted the two-stream-based framework [20]. In this architecture, a spatial stream processes individual RGB frames via a CNN, while a temporal stream captures motion information by stacking and processing optical flow images with another CNN. Finally, the classification scores of each stream are combined by a late fusion block. Contemporary methods, such as those proposed by the works of Molchanov et al. [17] and Kopuklu et al. [13], have adeptly integrated both RGB and OF data, resulting in significant improvements in recognition accuracy and a richer representation of hand gestures. However, despite the HGR advancements, a notable gap in the existing literature is the limited attention to spatial localization of hands, which is crucial when distinguishing multiple hands in a scene or determining the precise location of a gesture.

An important issue for the RGB+OF fusion is determining the optimal level within CNN models for information integration. Numerous efforts have fused multimodal information at different levels, namely early, late, and middle fusion [1,7,19]. Early fusion combines features at the data level or the early stages of the HGR architecture, while late fusion merges predictions from individual models as the last step of the architecture. Moreover, some methods have shown that mid-fusion is more effective because it captures intricate correlations between modalities throughout the entire network [7,12,19]. The middle multi-level fusion strategy has notable contributions, such as those by Joze et al. [12] and recently Hampiholi et al. [7], emphasizing the potential of middle fusion when integrating RGB and OF data. However, many existing approaches often neglect challenges like architectural complexity, inefficiencies from traditional fusion blocks, and the computational costs introduced by OF estimators. This highlights the urgency for methodologies that seamlessly balance performance and computational efficiency, the core objective of this work.

3 Proposed Method

In this section, we detail our YOLO-based single-stage architecture that leverages the benefits of both RGB and OF data to achieve robust and efficient Dynamic Hand Gesture Recognition and Localization (HGRL). Figure 2 presents a block diagram detailing the entire architecture, capturing the flow from the multimodal RGB+OF input to the precise HGRL output. The diagram offers a holistic perspective of our architecture. However, for a more detailed understanding, we'll dissect it into its core segments: the Backbone, the RGB+OF fusion, the Neck, and the Head. These elements are crucial, each contributing uniquely to the balance between speed and accuracy in our method.

Backbone. The backbone serves as the foundational structure of the architecture, responsible for initial feature extraction. Our design adopts the CSP-Darknet53 structure, a modification of the renowned Darknet architecture [18] based on the CSP (Cross Stage Partial) principle [24]. This structure efficiently enhances learning by decomposing the feature map from the previous stage into two parts and then merging them with a convolution layer after a series of bottlenecks, as illustrated in the CSP$_N$-2D diagram of Fig. 2. Note that the output size of feature maps is specified in each block of the diagram. The core of the backbone is constructed by stacking multiple 2DConv-CSP modules, with each 2DConv layer comprising a 3×3 Convolution, followed by Batch Normalization, and then activated by the SiLU (Sigmoid Linear Unit) function [4]. In summary, the backbone effectively captures both low-level and high-level features, setting a robust foundation for the subsequent processing stages.

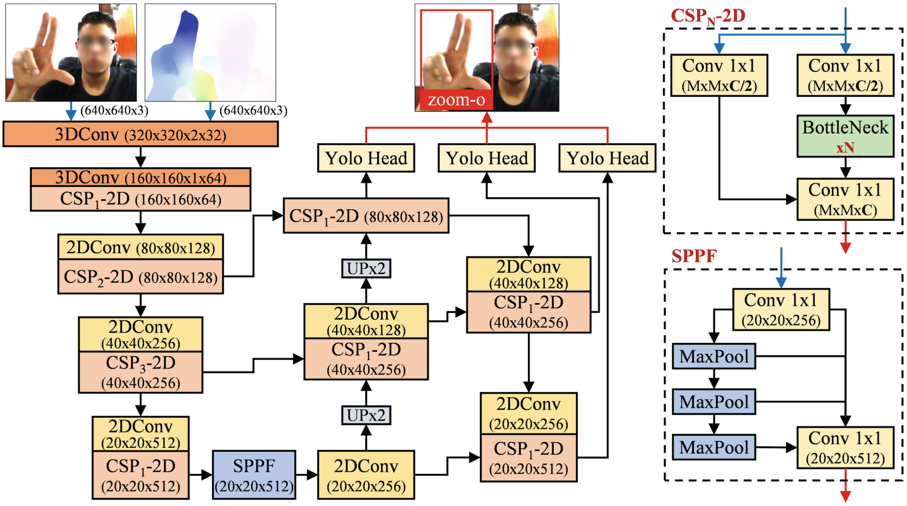

Fig. 2. Block diagram of our YOLO-based architecture with efficient 3Dconv fusion of RGB and OF for HGRL.

RGB+OF Fusion. Crucial to our architecture is the fusion of RGB and OF data. These modalities, when combined, offer a comprehensive view of dynamic hand gestures comprising appearance (RGB) and motion dynamics (OF). Our chosen methodology for fusion leverages 3Dconv layers, designed to extract spatiotemporal features by processing depth-wise spatial information across consecutive time steps. This is particularly advantageous for our input, which consists of two consecutive images (RGB and OF) of size 640×640×3, as depicted in Fig. 2. Thus, the input is initially shaped as 640×640×2×3. It then undergoes two 3Dconv layers as follows:

- The data first passes through a 3Dconv layer with 32 filters and a spatial stride of 2. This step processes the RGB and OF frames together, yielding an output shape of 320×320×2×32.
- It then traverses another 3Dconv layer with 64 filters and spatial and temporal strides of 2. This operation condenses the temporal dimension by effectively merging the RGB and OF data, resulting in a 160×160×1×64 feature map shape.
- The rest of the 2DCNN-based architecture processes this fused data by simply reshaping the feature maps to 160×160×64.

In this way, our fusion strategy doesn't just append one modality to another. Instead, it learns and retains the spatial and temporal intricacies of both modalities, setting the stage for subsequent processing stages to work with a richer feature representation.

Neck. Acting as an essential conduit between the foundational Backbone and the decisive Head, the Neck in our architecture ensures efficient high-to-low level feature communication. We incorporate the SPPF (Spatial Pyramid Pooling Faster) structure [11], a faster variant of the traditional SPP [8]. SPPF enables the network to capture multi-scale features by pooling feature maps at different scales and concatenating them, as shown in Fig. 2, resulting in a robust representation resistant to varying object sizes. Alongside, we employ the CSP-PAN (CSP - Path Aggregation Network) structure, a modification of the conventional PAN [14] built on the CSP principle [24]. CSP-PAN effectively redistributes and aggregates feature maps across layers, enhancing the learning capability by enabling efficient cross-scale feature communication. In essence, the Neck ensures that fine-grained and coarse features are seamlessly bridged and prepared for the final detection phase.

Head. The Head of our architecture is the last stage, where the processed features are converted into predictions. Our implementation capitalizes on the YOLOv3 Head [18], known for its high efficiency in real-time object detection tasks. YOLOv3 introduces three sizes of anchor boxes at three different scales, catering to varying object sizes. Each of these scales uses its set of anchor boxes to predict both the bounding box coordinates and the objectness score. Additionally, each bounding box prediction predicts the confidence score, which signifies the probability of an object being present and how well the bounding box fits the object. As illustrated in Fig. 2, the Head also benefits from a three-tier detection mechanism, allowing detections at three different resolutions, thereby enhancing accuracy across a range of object sizes. This meticulous design ensures precise localization of hand gestures while optimizing computational overhead, making it particularly apt for our purpose.

In summary, our single-stage architecture begins with the robust CSP-Darknet53 backbone for feature extraction, then integrates RGB+OF inputs through a simple yet effective 3Dconv fusion process. This harmonized information then traverses the efficient SPPF and the adaptable CSP-PAN in the Neck, culminating in the precise and real-time detection capabilities of the YOLOv3 Head. As illustrated in Fig. 2, our design prioritizes both accuracy and computational efficiency, presenting a significant contribution to HGRL.

4 State-of-the-art Optical Flow Methods

Dense Optical Flow (OF) consists of estimating pixel motion between consecutive frames in a video sequence, making it an invaluable asset, particularly for HGRL applications. As the movement of hands and fingers becomes intricate and fast-paced, reliable OF estimation becomes essential to distinguishing gestures accurately. Despite several OF methods existing in the literature, we focus our selection on approaches that have demonstrated state-of-the-art (SOTA) performance on conventional benchmarks, such as Sintel [3] and KITTI [5]. This section explores five leading methods, highlighting their contributions and evaluating their computational and performance capabilities.

RAFT: Recurrent All-Pairs Field Transforms [22]**.** Deviating from conventional approaches that compute per-pixel displacements, RAFT employs recurrent neural networks and constructs a 4D cost volume to analyze all possible pixel pairs, ensuring a more cohesive motion estimate across frames. Utilizing ConvGRU, it iteratively updates a dense flow field, beginning with a coarse level and progressively refining it. This guarantees exceptional precision even in complex dynamic scenes. This methodology not only contributes to its superior accuracy but also results in visually coherent and smooth flow fields. RAFT has been validated extensively, being the backbone of several incremental improvements, underscoring its efficacy and relevance in the field of optical flow estimation.

GMA: Global Motion Aggregation [10]**.** GMA is a novel approach that specifically targets the intricate issues presented by occlusions when estimating OF. The authors introduce a transformer-based methodology that leverages an attention mechanism to identify long-range interdependencies between individual pixels in the reference frame. This attention mechanism, often referred to as the GMA block, enables the model to give varying importance to different pixels, ensuring a more accurate and nuanced estimation, especially in occluded regions. Instead of solely relying on local evidence, the method aggregates the motion characteristics of these pixels on a global scale. This holistic approach results in a more accurate and detailed representation of motion, particularly in regions with prevalent occlusions. By incorporating GMA features into the RAFT framework [22], a new SOTA performance has been established.

KPA: Kernel Patch Attention [15]**.** Despite the considerable advancements made by deep learning-based OF methods, their primary emphasis lies in learning and measuring feature similarities, often neglecting the spatial relations that reveal motion affinities. KPA addresses this by introducing kernel patch attention, which operates on each local patch to determine context affinities for better flow field inference. Traditional optical flow algorithms emphasized both feature similarities and spatial smoothness. In contrast, KPA effectively blends both aspects, focusing on local relations based on context and spatial affinities. The proposed KPA operator employs a patch-based sliding window strategy, offering a comprehensive solution for reliable motion understanding. Once more, the KPA method builds upon the RAFT framework [22] to achieve a new SOTA performance on standard benchmarks.

SKF: Super Kernel Flow Network [21]**.** Similar to GMA, the SKF method is proposed to mitigate the impacts of occlusions in OF estimation. This approach benefits from super kernels (SK), which provide enlarged receptive fields to complement absent matching information and recover occluded motions. SKF introduces an efficient architecture utilizing a conical design with residual connections, which splits the convolution operation into depth-wise convolutions, consisting of a large depth-wise kernel and an auxiliary smaller depth-wise kernel. SKF introduces an efficient architecture with a conical design complemented by residual connections. This design aims to split the convolution operation into depth-wise convolutions, using both a large depth-wise kernel and an auxiliary smaller depth-wise kernel. Despite the architecture reminiscence of GMA, especially in using the GMA module, SKF distinguishes itself through the innovative application of SK modules. It also differs from the RAFT framework by utilizing the SK block as an updater instead of ConvGRU.

FlowFormer: Transformer Architecture for OF [9]**.** FlowFormer presents a novel approach representing a fusion of transformer architectures with established OF estimation techniques. It takes inspiration from the rising popularity of transformers, known for modeling long-range relations. Unlike directly operating on image pixels that demand a large number of parameters and training samples, FlowFormer incorporates the advantages of the cost volume from previous techniques. It employs an encoder-decoder architecture that transforms the 4D cost volume into compact, globally aware latent cost tokens. The proposed cost decoder also adopts a recurrent attention layer inspired by RAFT [22]. This decoder treats cost decoding as a recurrent query process with dynamic positional cost queries, delivering state-of-the-art performance.

Comparative Analysis. Table 1 presents a detailed comparison of the performance and computational costs of the five OF methods. On evaluating the benchmarks, the most recent addition to the field (FlowFormer, ECCV'22) exhibits superior accuracy on the Sintel dataset, registering the lowest error rate of 2.09. However, the KPA establishes superiority on the KITTI dataset with an error rate of 4.60. Regarding computational efficiency, RAFT stands out for both input resolutions, demanding 242.8 and 60.7 GFLOPs for 640×480 and 320×240 resolutions, respectively. FlowFormer, despite its leading performance, demands a computational burden nearly three times heavier than that of RAFT. Nonetheless, GMA and SKF present a good trade-off of performance and efficiency. In the next section, we delve deeper, evaluating the significance of these findings in the RGB+OF context for HGRL.

Table 1. Computational cost and performance on standard benchmarking of the five analyzed OF methods.

		GFLOPs		Results	
Method	Params	(640×480)	(320×240)	Sintel	KITTI
RAFT (ECCV'20)	5.26M	242.8	60.7	2.86	5.10
GMA (ICCV'21)	5.88M	272.3	68.1	2.47	4.93
KPA (CVPR'22)	5.99M	327.6	82.2	2.36	**4.60**
SKF (NeurIPS'22)	6.27M	295.2	73.8	2.27	4.84
FlowFormer (ECCV'22)	16.17M	756.5	173.5	**2.09**	4.68

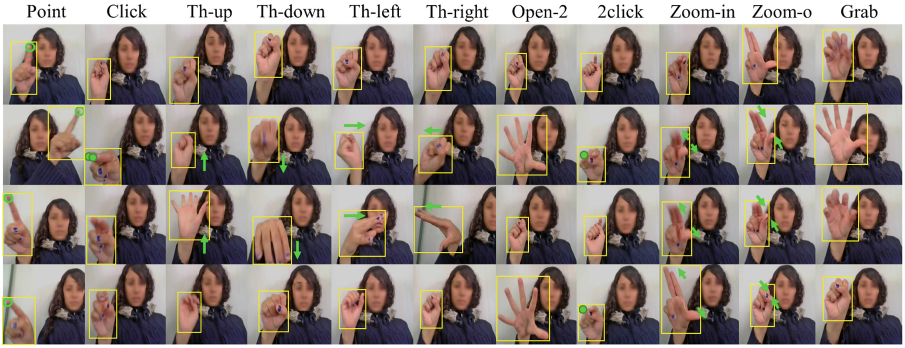

Fig. 3. Dynamic gestures with hand annotations from the IPN Hand dataset used in the HGRL Evaluation.

5 Experimental Results

5.1 Dataset

In this paper, we utilize the IPN Hand dataset [2], a comprehensive collection of dynamic gestures tailored for touchless screen interaction. The dataset comprises RGB videos recorded at 640×480 resolution and 30fps using PC or laptop cameras. The videos originate from 28 distinct scenes involving 50 participants, including challenges such as cluttered backgrounds and varying illumination conditions.

For our evaluation, we assess the performance of HGRL on 11 specific gestures, illustrated in Fig. 3, which account for a total of 3,457 gesture instances. Given the absence of hand location data in the dataset, we manually annotate 82,769 keyframes from all instances. To facilitate this, we trained a YOLOv5 [11] model on a smaller dataset to produce candidate hand locations. Subsequently, we manually refine the hand annotations for each instance at an approximate rate of 9 fps. The training set consists of 2,531 gesture instances from 37 subjects, translating to 64,768 annotated frames. Conversely, the test set includes 926 instances from 13 subjects, generating 18,001 annotated frames.

5.2 Implementation Details

We use Python 3.7.16 and PyTorch 1.10.2 with CUDA 12.0 on an Intel Core i7-9700K desktop with a single Nvidia GTX 2080Ti GPU for all experiments. To train the proposed YOLO-based architecture with RGB+OF 3Dconv fusion, we set 30 epochs using a batch size of 32 and cropped regions of size 640×640. We initialized the CSP-Darknet53 backbone with pre-trained weights from ImageNet, specifically for the 2DConv layers, while training all other layers from scratch. The optimization approach was Stochastic Gradient Descent, with a momentum of 0.937, a learning rate of 0.01, and a weight decay of 0.0005. The loss functions utilized were the Binary Cross Entropy (BCE) for class and objectness evaluations, and the Complete Intersection over Union (CIoU) for location loss, as in [11]. Furthermore, we incorporated Mosaic Augmentation alongside random rotation, scaling, and translation as part of our data augmentation strategy.

For OF approaches, we leveraged the official open-source implementations and pre-trained models released by the authors of each method. We obtained OF representations from the 82,769 annotated frames of the IPN hand dataset.

5.3 Analysis of of in the Proposed Fusion Framework

In this section, we delve into the impact of OF integration within our RGB+OF framework. A comprehensive analysis was performed to objectively assess each OF method's contribution to our fusion scheme. Our evaluation is based on standard metrics, such as Precision, Recall, and the mean Average Precision (mAP) at varying Intersection over Union (IoU) thresholds.

Table 2 presents the results for our RGB+OF fusion model with different SOTA Optical Flow methods, benchmarked against the "RGB only" results. As expected, the fusion models consistently surpass the RGB-only metrics. In particular, the RGB+SKF combination achieves the highest scores in all categories, yielding an average 10% improvement over the baseline. This emphasizes SKF's ability to represent motion nuances, enhancing gesture recognition when combined with RGB. These findings reinforce that integrating OF can significantly augment gesture recognition performance, primarily when implemented with the right OF methodology.

Table 2. Evaluation of the proposed RGB+OF fusion model with SOTA OF methods.

Method	Precision	Recall	mAP@0.5	mAP@0.5:0.95
RGB-only	*54.24*	*61.23*	*57.85*	*46.15*
RGB+RAFT	56.59	69.95	64.66	57.47
RGB+GMA	56.11	70.07	64.84	57.88
RGB+KPA	56.88	68.27	65.25	58.45
RGB+SKF	**58.04**	**71.61**	**66.98**	**59.47**
RGB+FlowFormer	56.98	68.07	64.40	57.77

For a more detailed analysis of the impact of OF integration, we present class-specific Average Precision (AP) results. As illustrated in Fig. 4, the RGB+SKF combination outperforms the RGB-only approach for most gesture classes. For example, the "Point" gesture, one of the fundamental human-computer interactions, witnesses a 5% increase in AP when augmented with SKF. Similarly, more complex gestures like "2click" and "Zoom-in" observe substantial improvements of more than 10% and 15%, respectively. However, for a couple of classes, such as "Th-down" and "Open-2", the RGB+SKF does not achieve top results.

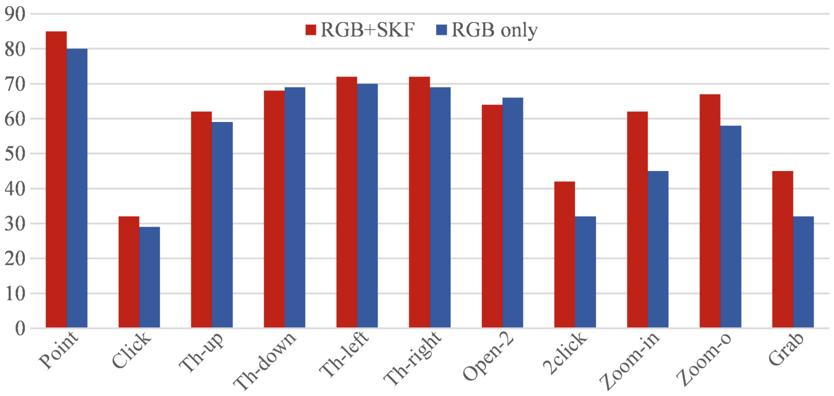

Fig. 4. Class-AP comparison between RGB+SKF and RGB-only methods.

Still, these differences are minimal, suggesting the overall positive impact of the RGB+OF fusion. This detailed analysis reinforces the importance of blending motion information, particularly when recognizing gestures with complex motions.

Figure 5 presents qualitative results of gesture recognition capabilities across different methods. The static nature of RGB-only makes it challenging to discern the gestures in the first three examples, making the integration of RGB+OF vital. Most RGB+OF approaches converge in their predictions, particularly for 'Zoom-o' and 'Zoom-in' gestures. Nevertheless, discrepancies arise in the third and fourth examples, highlighted by RGB-only's 'Grab' misclassification and the spurious gestures detected by GMA and FlowFormer due to the user's head movements. Interestingly, FlowFormer's representation in the occluded-hand scenario of the first example leans towards image appearance rather than actual motion. This figure highlights the importance of accurately detecting motion to improve gesture recognition in RGB+OF methods.

5.4 Ablation Study

Finally, we conducted an ablation study to explore the effectiveness of different fusion methods in our RGB+OF framework. Table 3 presents the comparative results, where the baseline RGB-only model serves as a reference. The initial attempt to integrate OF using 2DConvs exhibited a drop in precision but an increase in the remaining metrics. However, the transition to 3DConvs displayed evident advantages. A single layer of 3DConv brought significant improvements in mAP over the RGB-only baseline, with negligible computational overhead. Our proposed method, which incorporates two layers of 3DConvs (3Dconv$_{2layers}$), achieves the highest recall and nearly the best mAP@0.5 with just a marginal increase in parameters and GFLOPs. Further, adding a CSP-3D block increased the precision but also added considerable computational burden, increasing the GFLOPs by 11.3 compared to our proposed

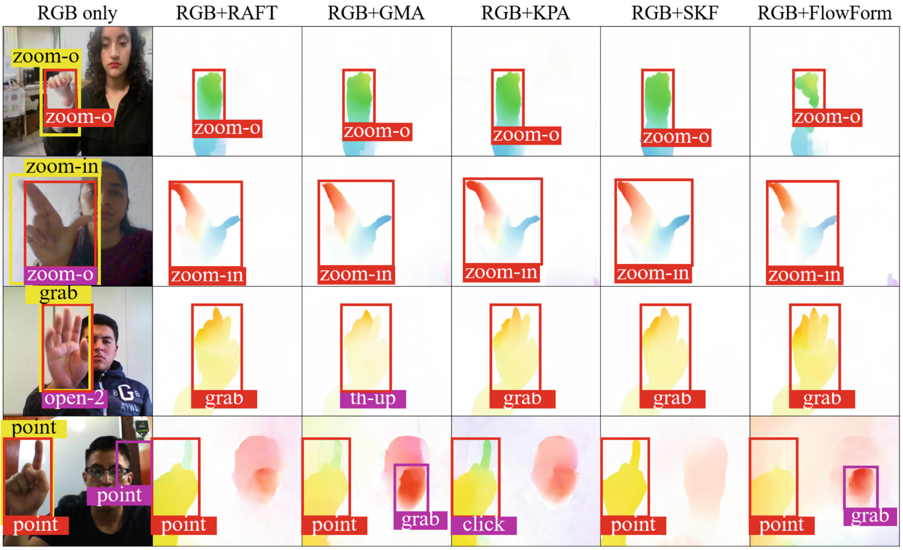

Fig. 5. Qualitative results of HGRL and the corresponding optical flow representations.eps

model. Notably, the increase in GFLOPs for our models is negligible when considering the computational requirements of optical flow estimators. This accentuates that most of the computation lies in the OF estimation and not in the fusion technique itself, making our choices in model design effective and efficient.

Table 3. Ablation study of different fusion approaches.

Method	Params	GFLOPs	Precision	Recall	mAP@0.5
RGB-only	*7.05*M	*16.1*	*54.24*	*61.23*	*57.85*
2Dconv	7.06M (+0.01)	16.8 (+0.7)	47.78	68.99	60.42
3Dconv$_{1layer}$	7.06M (+0.01)	16.8 (+0.7)	57.11	69.83	65.49
3Dconv$_{2layers}$	7.10M (+0.05)	17.2 (+1.1)	58.04	**71.61**	66.98
3D$_{2layers+CSP-3D}$	7.24M (+0.19)	27.4 (+11.3)	**58.36**	71.36	**67.02**

6 Conclusion

In this paper, we explored the integration of Optical Flow (OF) into RGB-based hand gesture recognition and localization. We found that our YOLO-based architecture with RGB+OF 3DConv fusion consistently surpassed the RGB-only baseline, especially when fused with SKF. Our ablation study highlighted the minimal computational overhead added by our fusion technique, emphasizing that the core computational cost lies in OF estimation. Qualitative results further illustrated challenges and potential areas for improvement. For future work, we aim to investigate loss functions that can penalize the RGB+OF contribution. Additionally, we intend to enrich temporal information by incorporating multiple Optical Flow representations.

References

1. Asadi-Aghbolaghi, M., et al.: A survey on deep learning based approaches for action and gesture recognition in image sequences. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 476–483. IEEE (2017)
2. Benitez-Garcia, G., Olivares-Mercado, J., Sanchez-Perez, G., Yanai, K.: IPN hand: a video dataset and benchmark for real-time continuous hand gesture recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 4340–4347. IEEE (2021)
3. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 611–625. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_44
4. Elfwing, S., Uchibe, E., Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Netw. **107**, 3–11 (2018)
5. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361. IEEE (2012)
6. Guo, L., Lu, Z., Yao, L.: Human-machine interaction sensing technology based on hand gesture recognition: a review. IEEE Trans. Hum.-Mach. Syst. **51**(4), 300–309 (2021)
7. Hampiholi, B., Jarvers, C., Mader, W., Neumann, H.: Convolutional transformer fusion blocks for multi-modal gesture recognition. IEEE Access **11**, 34094–34103 (2023)
8. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
9. Huang, Z., et al.: FlowFormer: a transformer architecture for optical flow. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. ECCV 2022. LNCS, vol. 13677. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19790-1_40
10. Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.: Learning to estimate hidden motions with global motion aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9772–9781 (2021)
11. Jocher, G., et al.: ultralytics/yolov5: v3.0. https://github.com/ultralytics/yolov5/. Accessed 15 Feb 2023
12. Joze, H.R.V., Shaban, A., Iuzzolino, M.L., Koishida, K.: MMTM: multimodal transfer module for CNN fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13289–13299 (2020)
13. Kopuklu, O., Kose, N., Rigoll, G.: Motion fused frames: data level fusion strategy for hand gesture recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2103–2111 (2018)

14. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768 (2018)
15. Luo, A., Yang, F., Li, X., Liu, S.: Learning optical flow with kernel patch attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8906–8915 (2022)
16. Molchanov, P., Gupta, S., Kim, K., Kautz, J.: Hand gesture recognition with 3D convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–7 (2015)
17. Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., Kautz, J.: Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4207–4215 (2016)
18. Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
19. Roitberg, A., Pollert, T., Haurilet, M., Martin, M., Stiefelhagen, R.: Analysis of deep fusion strategies for multi-modal gesture recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
20. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
21. Sun, S., Chen, Y., Zhu, Y., Guo, G., Li, G.: SKFlow: learning optical flow with super kernels. Adv. Neural. Inf. Process. Syst. **35**, 11313–11326 (2022)
22. Teed, Z., Deng, J.: RAFT: recurrent all-pairs field transforms for optical flow. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 402–419. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_24
23. Wachs, J.P., Kölsch, M., Stern, H., Edan, Y.: Vision-based hand-gesture applications. Commun. ACM **54**(2), 60–71 (2011)
24. Wang, C.Y., Liao, H.Y.M., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H.: CSPNet: a new backbone that can enhance learning capability of CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 390–391 (2020)
25. Zengeler, N., Kopinski, T., Handmann, U.: Hand gesture recognition in automotive human-machine interaction using depth cameras. Sensors **19**(1), 59 (2019)
26. Zhang, W., Wang, J., Lan, F.: Dynamic hand gesture recognition based on short-term sampling neural networks. IEEE/CAA J. Automatica Sin. **8**(1), 110–120 (2020)