

# A Comparative Analysis of Decision-Level Fusion for Multimodal Driver Behaviour Understanding

Alina Roitberg      Kunyu Peng      Zdravko Marinov  
Constantin Seibold      David Schneider      Rainer Stiefelhagen

Institute for Anthropomatics and Robotics  
Karlsruhe Institute of Technology  
{firstname.lastname}@kit.edu

**Abstract**—Visual recognition inside the vehicle cabin leads to safer driving and more intuitive human-vehicle interaction but such systems face substantial obstacles as they need to capture different granularities of driver behaviour while dealing with highly limited body visibility and changing illumination. *Multimodal* recognition mitigates a number of such issues: prediction outcomes of different sensors complement each other due to different modality-specific strengths and weaknesses. While several late fusion methods have been considered in previously published frameworks, they constantly feature different architecture backbones and building blocks making it very hard to isolate the role of the chosen late fusion strategy itself.

This paper presents an empirical evaluation of different paradigms for decision-level late fusion in video-based driver observation. We compare seven different mechanisms for joining the results of single-modal classifiers which have been both popular, (e.g. score averaging) and not yet considered (e.g. rank-level fusion) in the context of driver observation evaluating them based on different criteria and benchmark settings. This is the first systematic study of strategies for fusing outcomes of multimodal predictors inside the vehicles, conducted with the goal to provide guidance for fusion scheme selection.

## I. INTRODUCTION AND RELATED WORK

*Multimodality* increasingly gains attention in driver observation systems [1], [2], [3], [4]: prediction outcomes of multiple sensors complement each other due to modality-specific strengths and weaknesses as well as different visibility (examples in Figures 1 and 2). Rising levels of automation increase human freedom, leading to drivers being engaged in distractive behaviours more often while the type of activities become increasingly diverse. This is very challenging for *unimodal* driver observation systems, which need to capture different complexities and granularities of situations inside the cabin despite strongly restricted body visibility. For example, frameworks developed for manual driving often focus on the face view to capture the attentiveness regarding the driving scene [3], [5], [6]. However, as the driver is gradually relieved from actively steering the car, activities such as *working on laptop* or *reading magazine*, which were almost unthinkable until now, become more common. Equipping the vehicle with multiple complementing sensors enables recognition of very different behaviour types, but *how to link the information* becomes an important question.

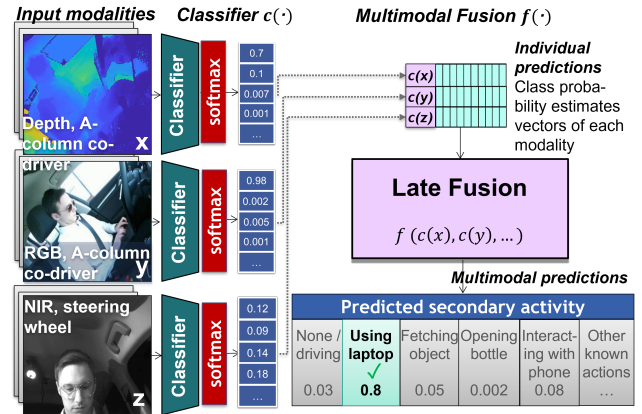


Fig. 1. A high-level overview of a multimodal driver observation framework featuring three separate classification streams, with their fusion is carried out after the single-modal predictions were obtained. We implement and study different techniques for linking such single-modal outcomes.

The state-of-the-art of multimodal driver activity recognition constantly changes depending on different architecture choices, losses and classifier components [7], [8], [9], [2], [1], [10], but a large portion of such methods employ late fusion via score averaging to link the information [2], [1], [9], [4]. Multimodal fusion algorithms can be grouped depending on the point of fusion, (e.g., early-, mid-, or late-fusion) and based on the methodology (learning- and decision-based approaches). The learning-based approaches *learn* to combine the streams (and can therefore be applied at different information processing stages). In *decision-level fusion*, on the other hand, individual unimodal probability estimates for each behaviour category are obtained a priori, after which a transformation function, such as average, product, or voting joins them into a common multimodal decision.

This work conducts the first systematic study of strategies for fusing outcomes of multimodal predictors at decision-level for visual recognition inside the vehicle cabin. Despite omitting intra-modality correlations at earlier stages, decision-level operations bring important advantages. First, in contrast to the learning-based methods, the multimodal systems operating on decision-level are highly modular, as the individual modalities with pretrained classifiers can be

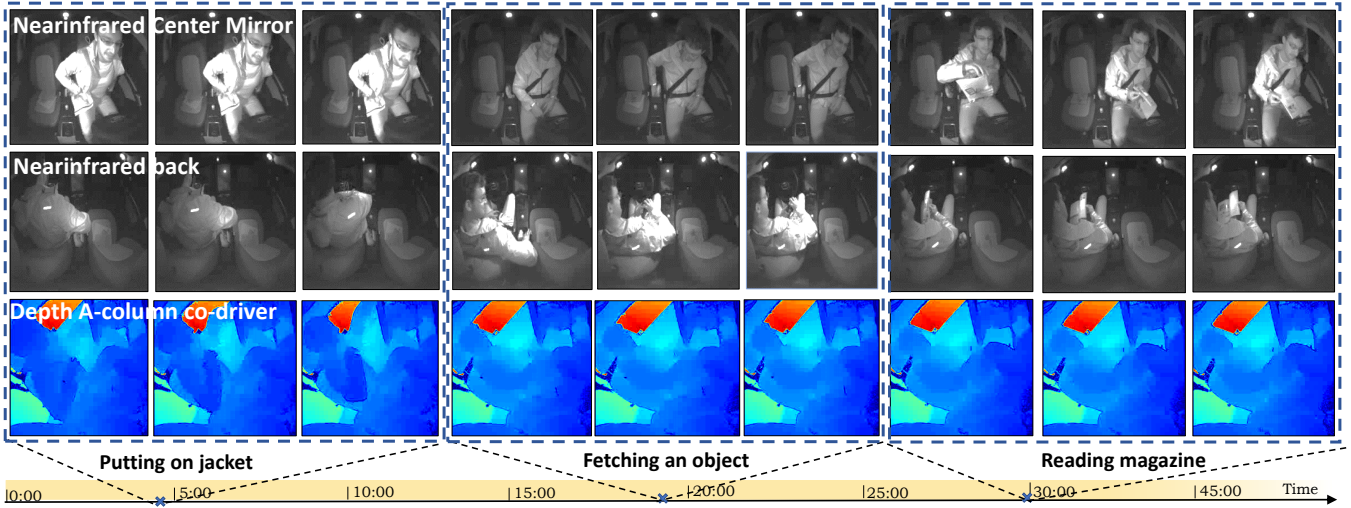


Fig. 2. Example of a multimodal driver activity recognition setting with highly distractive behaviours during automated driving. Different modalities have their specific strengths and limitations depending on the visible portion of the cabin and sensor-specific characteristics. For example, both RGB and NIR cameras might capture unnecessary textures constituting additional noise, while RGB sensors depend on the illumination. Depth data is less sensitive to illumination changes and skips unnecessary texture details (e.g., clothing) but might also miss details important for the behaviours-of-interest.

flexibly plugged-in or removed without any additional re-training. As a consequence, if one of the sensors is damaged, a decision-level fusion system would simply exclude it from contributing, whereas training of the fusion model would need to be revisited in standard learning-based approaches, as the feature vector appearance changes.

Among decision-level fusion techniques, averaging of the obtained *Softmax* scores is presumably the most common choice in driver activity recognition [2], [1], [9], [4]. In the broader fields of general machine learning and computer vision, this approach is also highly popular [11], [12], [13], [14], [15], [16], [17], [18], but other strategies, which were rather overlooked in the field of driver observation, such as the max rule [19], [14], [17], [20] or the product rule [12], [21], [19], [22], [23], [24], [25], [17], [20] also gained attention. A theoretical study of such methods from the pre-Deep Learning era is provided in [26]. Rank-level decision-level fusion, such as Borda Count voting [27], [28], [29] and Reciprocal Rank Voting [30] are less popular, but have been successfully applied in the field of biometric identification [31], [32]. There are several works targeting multimodal fusion through learning-based methods, *e.g.*, using SVM, LSTM, or neural network fusion layers [33], [34], [35], [7], [36], [37]. We, however, consider these out of the scope of this work, as these require additional training and cannot be directly used out-of-the-box for target fusion at the decision-level. Nevertheless, recent computer vision research is rather focused on the generation of high performing single-modal classifiers while fusion strategies are considered of lesser importance and few of them are systematically explored in combination with novel CNN-based methods.

**Summary and contributions** In this work, our goal is to implement and systematically evaluate different strategies for decision-level fusion in the context of multimodal driver behaviour assessment. We build upon recent advances in

driver observation and train a neural network often utilized in this task separately for each of the eight modalities of a standard multimodal driver activity recognition testbed [2]. We compare 10 different mechanisms for joining the results of single-modal classifiers which have been both popular, (*e.g.*, score averaging) and not yet considered, (*e.g.*, rank-level fusion) in the context of driver observation and evaluate them based on different criteria and benchmark settings. Our results indicate that the choice of fusion mechanisms impacts the model performance. Furthermore, the commonly employed average-fusion being outperformed by several other methods in all evaluation settings and metrics. Of the considered methods, product-fusion and max-fusion yielded the best recognition results. Interestingly, while max-fusion oftentimes outperformed product-fusion by a small margin, product-fusion is consistently more effective when it comes to top-5 accuracy, indicating, that it might be useful in coarser recognition. We further compare our multimodal system to the best performing unimodal view. Overall, multimodality is clearly beneficial for almost all behaviour types, but the effect depends on visibility and recognition difficulty: the largest benefits of multimodality were observed in driver behaviours with medium recognition difficulty. To the best of our knowledge, this is the first systematic study of strategies for decision-level fusion inside the vehicle cabin. Our experiments provide empirical evidence that the commonly employed late fusion via averaging is not the most effective way of linking unimodal driver observation results, and we hope that our study will provide guidance for better fusion scheme selection in the future.

## II. REVISITING LATE FUSION FOR VIDEO-BASED DRIVER OBSERVATION

In this paper, we analyze different approaches for fusing the decision-level predictions of multiple visual driver observation models. That is, given  $N$  different modalities with

inputs  $x_i, i \in 1 \dots N$  (see examples in Figure 2) and  $N$  pre-trained unimodal classifiers with predictions  $c_i(x_i), i \in 1 \dots N$  containing probability estimates for each category, our goal is to correctly identify the potentially distractive behaviour of the driver by linking the information of these different modalities effectively. To this intent, we employ the I3D architecture [16] as the unimodal classifiers backbone, which has shown excellent results in driver activity recognition [2], [38]. We train the models for each modality individually. Afterwards, we utilize different variants of the decision-level fusion module which takes multiple class probability estimates produced by the individual classifiers as input and joins them to reach the final multi-modal decision. Note, that we specifically target *decision-level* approaches that do not require any architecture training or changes in architecture. While multiple introduced approaches address multimodal fusion with learning-based methods [33], [34], [35], [7], such approaches are out of the scope of this work. In total, we implement seven different strategies for multimodal decision-level fusion, which we now discuss in detail.

#### A. Score-level fusion

In score-level fusion, the goal is to combine the predictions of  $N$  classifiers on a  $d$ -classification task based on their class probability estimates  $c(x_i)$ , where  $i \in \{1 \dots N\}$ . We investigate fusing the predictions  $c(x_i)$  via summation or averaging, maximum, and product of the probability vectors. For this, we introduce the following notation in Table I:

$N \in \mathbb{N}$	Number of classifiers.
$d \in \mathbb{N}$	Number of classes.
$c(x_i) \in \mathbb{R}^d$	Probability estimates of $i^{\text{th}}$ classifier.
$c(x_i)_j \in \mathbb{R}$	Probability estimate of $i^{\text{th}}$ classifier for $j^{\text{th}}$ class.
$c(X) := \{c(x_1) \dots c(x_N)\} \in \mathbb{R}^{d \times N}$	Set of all probability estimates.
$c(X)_j := [c(x_1)_j \dots c(x_N)_j] \in \mathbb{R}^N$	Predictions for $j^{\text{th}}$ class from all classifiers.
$r_{ij}$	Rank of $j^{\text{th}}$ class in $c(x_i)$ .

TABLE I  
NOTATION FOR ALL THE LATE FUSION EQUATIONS.

Note that the fusion results from all the methods we investigate can be used in combination with  $\text{argmax}(\cdot)$  to produce the final class prediction.

**Sum-fusion and score averaging:** The sum-fusion (often referred to as average-fusion)  $f_{SUM}(\cdot)$  for  $N$  classifiers is defined as:

$$f_{SUM}(c(X)) = \frac{1}{N} \sum_{i=1}^N c(x_i) \quad (1)$$

Note that the division by  $N$  does not change the ranking of the summed predictions, but serves to regularize the output to sum up to 1. This fusion strategies has presumably been the most popular choice for fusion at decision-level in driver observation [2], [1], [9], [4].

**Median-based fusion:** The median-fusion  $f_{MED}(\cdot)$  for  $N$  classifiers is defined as:

$$f_{MED}(c(X)) = [\text{med}(c(X)_1) \dots \text{med}(c(X)_d)] \quad (2)$$

where

$$\text{med}(x) = \begin{cases} \hat{x}_{(d+1)/2}, & \text{if } d \text{ is odd} \\ \frac{1}{2}(\hat{x}_{(d/2)} + \hat{x}_{(d/2)+1}), & \text{otherwise} \end{cases} \quad (3)$$

Here  $\hat{x}_d$  is defined as the  $d^{\text{th}}$  element of  $\hat{x}$  and  $\hat{x}$  is  $x$  sorted in ascending order.

**Max-fusion:** The max-fusion  $f_{MAX}(\cdot)$  for  $N$  classifiers is defined as:

$$f_{MAX}(c(X)) = [\max(c(X)_1) \dots \max(c(X)_d)] \quad (4)$$

**Product-fusion:** The product-fusion  $f_{PROD}(\cdot)$  for  $N$  classifiers is defined as:

$$f_{PROD}(c(X)) = \gamma \prod_{i=1}^N c(x_i) \quad (5)$$

where  $\gamma \in \mathbb{R}$  is used as a regularization of the output [39].

**Weighted sum- and product-fusion:** Inspired by recent progress of weighted pooling functions [40], we further implement variants of sum- and product-fusion, where the individual predictions are weighted via *Softmax*-normalization amplifying the contribution of the most certain class predictions. The weighted sum-fusion  $f_{WSUM}(\cdot)$  and weighted product-fusion  $f_{WPROD}(\cdot)$  for  $N$  classifiers are defined as:

$$f_{WSUM}(c(X)) = \frac{1}{N} \sum_{i=1}^N w_i c(x_i) \quad f_{WPROD}(c(X)) = \gamma \prod_{i=1}^N w_i c(x_i), \quad (6)$$

$$\text{where } w_i = \frac{e^{c(x_i)}}{\sum_{j=1}^N e^{c(x_j)}}, \quad i \in 1 \dots N.$$

#### B. Rank-level fusion

In contrast to score-level fusion, rank-level fusion leverages the class rankings of multiple classifiers. The magnitude of each class score plays a role only in the ordering of the classes into a ranking list for each classifier. We investigate Majority Voting, the original and weighted Borda Count, as well as Reciprocal Rank Fusion as strategies in this category.

**Majority Voting:** Majority voting first estimates the top-1 predicted behaviour for each individual modality, after which the category, which was predicted by the most unimodal classifiers is selected as the final decision. Let  $\text{pred}_i := \text{argmax}(c(x_i)) \in \mathbb{N}$  be the predicted class from the  $i^{\text{th}}$  classifier. The number of the top-1 predictions from all classifiers for class  $j$  would then be:

$$\#j = \#\{\text{pred}_i == j | i \in \{1 \dots N\}\} \quad (7)$$

where  $\#\{\cdot\}$  denotes the set cardinality. The majority voting  $\text{mv}(\cdot)$  for  $N$  classifiers is defined as:

$$\text{mv}(c(X)) = [\#\{1 \dots \#d\}] \quad (8)$$

**Borda Count:** Another way for combining predictions via late fusion is utilizing a voting system, such as Borda Count [27]. The Borda Count voting system is described algorithmically in Algorithm 1. The class probabilities  $c(x_i)$

	Fusion Method	#Mod=2				#Mod=4				#Mod=8			
		Balanced Acc.		Unbalanced Acc.		Balanced Acc.		Unbalanced Acc.		Balanced Acc.		Unbalanced Acc.	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Score-level	Avg. w/o weight. (standard)	47.81	70.49	42.57	67.71	51.44	77.88	46.06	75.41	54.69	80.9	49.72	78.72
	Average w. weight.	47.52	70.49	42.2	67.71	51.46	77.88	46.24	75.41	54.96	80.75	50.09	78.72
	Median	47.81	70.49	42.57	67.71	51.87	80.07	46.79	78.35	54.01	84.62	49.54	81.83
	Max	47.52	70.19	42.2	66.97	<b>53.26</b>	77.45	<b>48.07</b>	74.68	<b>55.96</b>	80.55	<b>50.64</b>	78.35
	Product w/o weight.	49.32	<b>74.98</b>	44.4	<b>72.66</b>	51.76	<b>83.41</b>	46.97	80.92	53.99	85.47	49.36	82.75
	Product w. weight.	<b>49.57</b>	74.84	<b>44.77</b>	72.48	51.76	<b>83.41</b>	46.97	<b>80.92</b>	53.85	85.47	49.17	82.75
Rank-level	Majority	47.81	70.49	42.57	67.71	51.98	77.62	46.42	75.23	54.75	80.66	49.91	78.53
	Borda count w/o. weight.	44.41	73.76	38.72	72.11	50.65	80.5	46.06	78.35	54.25	<b>85.91</b>	50.09	<b>83.49</b>
	Borda count w. weight.	47.81	70.62	42.57	67.34	51.51	77.6	46.06	75.05	54.53	81.06	49.54	79.27
	Reciprocal Rank	42.65	69.96	37.06	66.79	48.45	81.45	43.3	79.27	52.58	83.76	48.26	80.73

TABLE II  
PERFORMANCE OF LATE-FUSION METHODS ON **RARE** CLASSES OF THE DRIVE&ACT TEST SET

from all the unimodal models are given as an input. The first loop goes over each of the  $N$  classifiers. Their predictions  $c(x_i)$  are sorted in descending order so that a ranking list  $I$  is created with their indices. In the second loop, the best class prediction for each classifier is given  $k$  points, the second-best  $k-1$  points, etc., where  $k$  is a hyperparameter. This is done for all classifiers, and in the end, these points are added up for the final scoring  $\hat{y}$ .

The Borda Count voting resembles a preferential voting system, in contrast to a majoritarian one. This incorporates the uncertainty of each of the separate models' predictions. In other words, if a model is uncertain about the correct class and ranks it as a second alternative, its prediction would contribute with  $k-1$  points for the correct class, instead of 0 points in the case of using a majority vote. However, this relies on the assumption that the classifiers are able to rank the ground truth in their top  $k$  predictions, i.e. are not weak.

**Data:** Probability Estimates:  $c(x_i) \in \mathbb{R}^d$ ,  $i \in \{1 \dots N\}$ , where  $N = \# \text{classifiers}$

**Result:** Fused Class Scores:  $\hat{y} \in \mathbb{N}^d$

```

 $\hat{y} \leftarrow [0 \dots 0];$ 
for  $i \in \{1 \dots N\}$  do
   $I \leftarrow \text{descending\_argsort}(c(x_i));$ 
  for  $j \in \{k \dots 1\}$  do
     $\hat{y}[I[k-j]] += j;$ 
  end
end
return  $\hat{y};$ 

```

**Algorithm 1:** Borda Count Voting Strategy

**Reciprocal Rank Fusion (RRF):** The  $RRF$ [30] for  $N$  classifiers is defined as:

$$RRF(c(X)) = [rrf(1) \dots rrf(d)] \quad (9)$$

where

$$rrf(j) = \sum_{i=1}^N \frac{1}{m + r_{ij}} \quad (10)$$

Cormack et al. [30] introduce the hyperparameter  $m \in \mathbb{N}$  and claim that it mitigates the impact of high rankings by outlier systems.

**Weighted Borda Count:** The WBC is an extension of the original algorithm, where the score of each voter is weighted by the corresponding weighting vector  $w \in \mathbb{R}^d$ . The WBC for  $N$  classifiers is defined as:

$$WBC(c(X)) = w \odot BC(c(X)) = w \odot \hat{y} \quad (11)$$

where  $\odot$  is the element-wise multiplication operator. The vector  $w$  can be computed by an arbitrary weighting function. In our experiments we use the mean softmax outputs, i.e.  $w = f_{SUM}(c(X))$ . We also considered computing the weights via Softmax-normalization over the modalities (as done in the weighted sum- and product-fusion) but observed a significant performance decline. The reasoning behind  $w$  is to enhance the contribution of the most certain class predictions in the fusion stage [41].

### III. EXPERIMENTAL RESULTS

#### A. Testbed

We chose the multimodal Drive&Act dataset [2] as our evaluation testbed as it provides a diverse set of driver behaviours recorded with eight synchronized sensors, therefore enabling a comprehensive study of fusion techniques with a large set of modalities. Drive&Act modalities include one RGB-, one depth-, and six Near-Infrared (NIR) views with 12 hours recorded in total. The videos are labeled with a hierarchical annotation scheme, where 34 fine-grained activities constitute the main evaluation level. We follow the original evaluation protocol comprising three splits into *training*, *validation* and *test* with no intersection of drivers (10, 2 and 3 people respectively).

The 34 fine-grained activity classes of Drive&Act are unbalanced: the number of examples per behaviour type ranges from 19 (*taking laptop from backpack*) to 2797 (*sitting still*). Since machine learning models rely strongly on the amount of training data, we report the performance separately for *common*, *rare*, and *all* categories, as suggested in [38]. We report the top-1 and top-5 accuracies under balanced and unbalanced conditions. For the balanced accuracy, the metric is computed individually for each class and the average over all 34 behaviours is reported. The unbalanced accuracy is the percentage of correctly recognized examples



	Fusion Method	#Mod=2				#Mod=4				#Mod=8			
		Balanced Acc.		Unbalanced Acc.		Balanced Acc.		Unbalanced Acc.		Balanced Acc.		Unbalanced Acc.	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Score-level	Avg. w/o weight. (standard)	72.68	92.23	77.41	94.59	80.12	95.13	84.9	96.70	82.01	96.60	86.27	97.54
	Average w. weight.	72.18	92.22	76.87	94.59	80.00	95.14	84.67	96.70	81.96	96.64	86.20	97.56
	Median	72.68	92.23	77.41	94.59	79.66	95.64	84.69	97.24	81.22	96.84	85.98	97.71
	Max	71.88	92.18	76.55	94.51	79.28	95.06	83.70	96.58	<b>82.76</b>	96.47	85.84	97.36
	Product w/o weight.	<b>74.51</b>	<b>94.60</b>	<b>80.06</b>	<b>96.34</b>	80.67	<b>96.54</b>	85.59	<b>97.67</b>	82.44	97.01	<b>86.86</b>	<b>97.92</b>
	Product w. weight.	74.47	<b>94.60</b>	80.05	<b>96.34</b>	<b>80.71</b>	<b>96.54</b>	<b>85.62</b>	<b>97.67</b>	82.44	96.99	<b>86.86</b>	97.90
Rank-level	Majority	72.64	92.23	77.32	94.59	79.70	95.13	84.62	96.70	81.51	96.51	86.05	97.44
	Borda count w/o. weight.	65.76	92.88	73.95	95.03	77.83	96.18	83.85	97.44	80.18	<b>97.17</b>	85.64	98.08
	Borda count w. weight.	72.48	92.43	77.23	94.66	80.11	95.14	84.92	96.68	81.99	96.56	86.30	97.51
	Reciprocal Rank	65.88	92.37	74.08	95.29	75.36	95.37	82.32	97.31	79.26	96.33	85.32	97.56

TABLE III  
PERFORMANCE OF LATE-FUSION METHODS ON COMMON CLASSES OF THE DRIVE&ACT TEST SET

	Fusion Method	#Mod=2				#Mod=4				#Mod=8			
		Balanced Acc.		Unbalanced Acc.		Balanced Acc.		Unbalanced Acc.		Balanced Acc.		Unbalanced Acc.	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Score-level	Avg. w/o weight. (standard)	60.25	81.36	74.31	92.19	65.78	86.50	81.45	94.81	68.35	88.75	83.01	95.87
	Average w. weight.	59.85	81.36	73.79	92.19	65.73	86.51	81.25	94.81	68.46	88.7	82.98	95.88
	Median	60.25	81.36	74.31	92.19	65.76	87.86	81.32	95.56	67.62	90.73	82.74	96.29
	Max	59.70	81.18	73.49	92.06	<b>66.27</b>	86.26	80.53	94.63	<b>69.36</b>	88.51	82.70	95.67
	Product w/o weight.	61.91	<b>84.79</b>	76.89	<b>94.23</b>	66.21	<b>89.97</b>	82.15	<b>96.18</b>	68.22	91.24	<b>83.52</b>	96.57
	Product w. weight.	<b>62.02</b>	84.72	<b>76.91</b>	94.22	66.23	<b>89.97</b>	<b>82.18</b>	<b>96.18</b>	68.14	91.23	83.50	96.55
Rank-level	Majority	60.23	81.36	74.23	92.19	65.84	86.37	81.22	94.79	68.13	88.59	82.84	95.75
	Borda count w/o weight.	55.08	83.32	70.81	92.99	64.24	88.34	80.48	95.74	67.22	<b>91.54</b>	82.48	<b>96.78</b>
	Borda count w. weight.	60.15	81.53	74.15	92.23	65.81	86.37	81.46	94.76	68.26	88.81	83.03	95.88
	Reciprocal Rank	54.26	81.17	70.78	92.75	61.91	88.41	78.85	95.70	65.92	90.04	82.02	96.06

TABLE IV  
PERFORMANCE OF LATE-FUSION METHODS ON ALL CLASSES OF THE DRIVE&ACT TEST SET

over the complete dataset, (*i.e.*, in unbalanced settings the underrepresented classes acquire a smaller weight). The additional top-5 accuracy is especially useful on Drive&Act since we might be interested in coarser recognition and dismiss mistakes caused by highly similar classes (such as *opening* and *closing bottle*).

We use  $k = 5$ ,  $m = 60$  and  $\gamma = 1$  for Borda Count and Reciprocal Rank Fusion and product fusion according to the previous literature [30], [39]. For training the eight unimodal classifiers, the initial I3D weights are initialized using the Kinetics dataset [16], as done in the original Drive&Act work [2] and then optimized for driver behaviour classification with stochastic gradient descent using the initial learning of 0.01 (decreased by a factor of 10 after 50 and 100 epochs), momentum of 0.9, weight decay of  $1e-7$  and mini-batch size of 8. During training, temporal data augmentation samples clips of 64 frames and spacial data augmentation computes random crops of size  $224 \times 224$ .

## B. Results

The main objective of our experiments is to determine the impact of fusion strategies for the probability estimates of multimodal predictors in the context of driver observation, where averaging has presumably been the most common choice for fusion at decision-level [2], [1], [9], [4]. Tables II, III and IV display balanced and unbalanced top-1 and top-5 accuracies for different fusion schemes and *rare*,

*common* and *all* driver behaviour categories respectively. In all settings, we consider 2, 4 and all 8 Drive&Act modalities (the 2 and 4 modalities were chosen by selecting the first 2/4 modalities from a random permutation of all available views). In Table II (underrepresented behaviours), product-fusion and max-fusion yielded the best outcome (for example, 1.82%, 5.53%, 2.01% and 5.51% gain in performance compared to the conventional score averaging for the different metrics and four modalities). Interestingly, the models with best results in terms of the top-1 accuracy are not necessarily the best as it comes to the top-5 results. This hints that some models are better at coarser recognition, since the top-5 metrics often omits fine-grained confusions, such as *preparing food* vs. *eating*. For instance, Borda Count is the best performing fusion method for 8 modalities in terms of the top-5 accuracy, while it usually yields similar or slightly worse results compared to averaging looking at the top-1 metrics. While additional weighting does not have a significant influence on product- and average-fusion, it positively impacts the Borda Count results.

These results are confirmed through our experiments on *common* and *all* categories (Tables III and IV): product- and max-fusion alternate in being the frontrunner, while averaging is not the most effective choice in all settings. Interestingly, while max-fusion oftentimes outperformed product-fusion by a small margin, product-fusion is consistently more effective as it comes to top-5 accuracy, indicating, that it

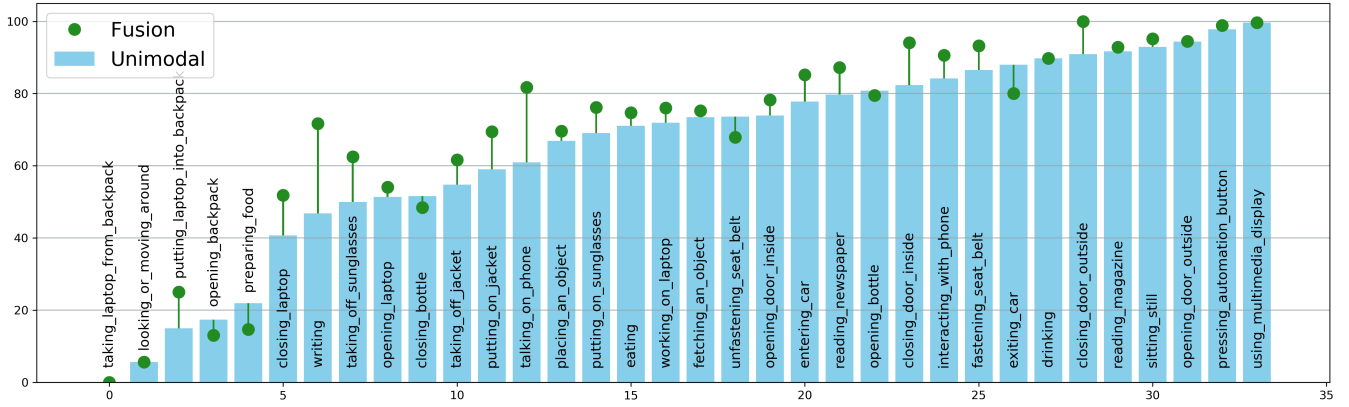


Fig. 3. Per-category accuracy for the best unimodal classifier (blue bar) and a multimodal model with eight views and product-fusion (green dot).

Modality	Balanced Acc.		Unbalanced Acc.	
	Top-1	Top-5	Top-1	Top-5
Center mirror, NIR	63.09	<b>88.60</b>	77.80	<b>94.63</b>
A-Column driver, NIR	59.92	87.19	73.69	94.09
Face view, NIR	42.32	70.23	55.74	84.84
Ceiling (back view), NIR	61.87	84.18	76.84	93.03
A-Column co-driver, NIR	<b>65.05</b>	87.52	<b>78.59</b>	94.38
A-Column co-driver, RGB	62.70	84.52	74.80	92.91
A-Column co-driver, Depth	59.83	84.41	71.73	92.47
Multimodal (product)	68.22	91.24	83.52	96.57

TABLE V

UNIMODAL PERFORMANCE FOR ALL CLASSES IN DRIVE&ACT

might be useful in coarser recognition. Overall, score-level approaches suit better than ranking-based strategies (with very few exceptions, where Borda Count is effective in terms of the top-5 accuracy).

As expected, utilizing more modalities positively impacts the recognition rates (for example, we achieve the top-1 balanced accuracy of 61.91%, 66.21% and 68.22% for 2, 4, and 8 modalities and all categories, see Table IV). As previously mentioned, the modality choice was conducted via a random permutation of all Drive&Act data sources. Since the first modality in the resulting sequence was *A column co-driver, depth*, adding 1, 3 and 7 additional modalities improves the unimodal performance by 2.1%, 6.38% and 8.39% accordingly (see Table V for the unimodal results). Lastly, in Figure 3 we compare our multimodal system (eight modalities with product-fusion) to the best performing unimodal view, which is *A column co-driver, NIR* according to Table V. The individual categories in Figure 3 are sorted by their accuracy in the unimodal setting, giving insight on how hard-to-recognize these behaviour types are. Overall, multimodality leads to performance improvement in almost all behaviour types, but the effect is different depending on the visibility and recognition difficulty: the largest benefits of multimodality were observed in driver behaviours with medium recognition difficulty. For instance, classification of examples with the driver *writing*, *taking off sunglasses* or *talking on phone* was improved by 24.86%,

12.5% and 20.81%. For “easier” driver behaviours, using more modalities positively influenced the performance but the effect is rather small (for example, only 2.24% improvement for sitting still). This is not surprising, as one effective modality might be already sufficient to recognize such activities. Interestingly, the results were rather mixed for very “hard to recognize” driver states, as the performance is improved in some cases (10% increase for *putting laptop into backpack* but 4% and 7% decline for *opening backpack* and *preparing food*, which is often confused with *eating*). Since we considered the best performing unimodal classifier, we believe that for certain difficult categories this modality was overwhelmingly better than other sensors, which rather constituted additional noise. The choice of modalities should therefore depend on the recognition use-case and behaviours-of-interest, but if a broad range of diverse secondary driver behaviours is required, multimodality is a powerful tool as it complements the advantages and unique characteristics of the individual sensors.

#### IV. CONCLUSION

In this work, we revisit the paradigm of decision-level fusion in the context of multimodal driver observation, where the predictions of the individual unimodal classifiers were predominantly joined via score averaging in the past [2], [1], [9], [4]. We operationalize and study different variants of seven decision-level fusion paradigms used in general machine learning literature in the context of driver behaviour understanding. We train eight unimodal classifiers on data provided by eight different cameras placed inside the vehicle cabin using a standard backbone neural network for driver activity categorization and equip them with different types of decision-level fusion modules for linking the probability estimates in a final decision. We found that late fusion based on the product-rule and max-rule lead to the best recognition results, but the effect depends on the task difficulty and number of modalities. This suggests that while the selection of the fusion scheme impacts the driver activity recognition performance noticeably, the conventional strategy of averaging the prediction scores is usually not the best choice.

**Acknowledgements** This work was partially supported by the Competence Center Karlsruhe for AI Systems Engineering (CC-KING) sponsored by the Ministry of Economic Affairs, Labour and Housing Baden-Württemberg.

## REFERENCES

- [1] O. Kopuklu, J. Zheng, H. Xu, and G. Rigoll, "Driver anomaly detection: A dataset and contrastive learning approach," in *IEEE/CVF Winter Conference on Applications of Computer Vision*.
- [2] M. Martin\*, A. Roitberg\*, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhausen, "Drive&Act: A Multi-modal Dataset for Fine-grained Driver Behavior Recognition in Autonomous Vehicles," in *ICCV*. IEEE, October 2019.
- [3] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena, "Car that knows before you do: Anticipating maneuvers via learning temporal driving models," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 3182–3190, 2015.
- [4] S. S. Khan, Z. Shen, H. Sun, A. Patel, and A. Abedi, "Modified supervised contrastive learning for detecting anomalous driving behaviours," *arXiv preprint arXiv:2109.04021*, 2021.
- [5] A. Rangesh, B. Zhang, and M. M. Trivedi, "Driver gaze estimation in the real world: Overcoming the eyeglass challenge," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1054–1059.
- [6] E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *Transactions on intelligent transportation systems*, vol. 15, no. 6, pp. 2368–2377, 2014.
- [7] N. Kose, O. Kopuklu, A. Unnervik, and G. Rigoll, "Real-time driver state monitoring using a cnn based spatio-temporal approach," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*.
- [8] J.-C. Chen, C.-Y. Lee, P.-Y. Huang, and C.-R. Lin, "Driver behavior analysis via two-stream deep convolutional neural network," *Applied Sciences*, vol. 10, no. 6, p. 1908, 2020.
- [9] M. Martin, J. Popp, M. Anneken, M. Voit, and R. Stiefelhausen, "Body pose and context information for driver secondary task detection," in *Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 2015–2021.
- [10] Z. Wharton, A. Behera, Y. Liu, and N. Bessis, "Coarse temporal attention network (cta-net) for driver's activity recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1279–1289.
- [11] J. Ye, K. Li, G.-J. Qi, and K. A. Hua, "Temporal order-preserving dynamic quantization for human action recognition from multimodal sensor streams," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015, pp. 99–106.
- [12] J. Imran and P. Kumar, "Human action recognition using rgb-d sensor and deep convolutional neural networks," in *2016 international conference on advances in computing, communications and informatics (ICACCI)*. IEEE, 2016, pp. 144–148.
- [13] F. Baradel, C. Wolf, and J. Mille, "Human action recognition: Pose-based attention draws focus to hands," in *IEEE International Conference on Computer Vision Workshops*, 2017, pp. 604–613.
- [14] S. Ardianto and H.-M. Hang, "Multi-view and multi-modal action recognition with learned fusion," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1601–1604.
- [15] N. Dawar and N. Kehtarnavaz, "A convolutional neural network-based sensor fusion system for monitoring transition movements in healthcare applications," in *2018 IEEE 14th International Conference on Control and Automation (ICCA)*. IEEE, 2018, pp. 482–485.
- [16] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.
- [17] C. Dhiman and D. K. Vishwakarma, "View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics," *IEEE Transactions on Image Processing*, 2020.
- [18] J. Cai, N. Jiang, X. Han, K. Jia, and J. Lu, "Jolo-gcn: mining joint-centered light-weight information for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2735–2744.
- [19] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng, "Deep convolutional neural networks for human action recognition using depth maps and postures," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 9, pp. 1806–1819, 2018.
- [20] S. S. Rani, G. A. Naidu, and V. U. Shree, "Kinematic joint descriptor and depth motion descriptor with convolutional neural networks for human action recognition," *Materials Today: Proceedings*, 2021.
- [21] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona, "Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 595–604.
- [22] P. Khaire, J. Imran, and P. Kumar, "Human activity recognition by fusion of rgb, depth, and skeletal data," in *Proceedings of 2nd International Conference on Computer Vision & Image Processing*, B. B. Chaudhuri, M. S. Kankanalli, and B. Raman, Eds. Singapore: Springer Singapore, 2018, pp. 409–421.
- [23] N. Dawar, S. Ostadabbas, and N. Kehtarnavaz, "Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition," *IEEE Sensors Letters*, vol. 3, no. 1, pp. 1–4, 2018.
- [24] C. Zhao, M. Chen, J. Zhao, Q. Wang, and Y. Shen, "3d behavior recognition based on multi-modal deep space-time learning," *Applied Sciences*, vol. 9, no. 4, p. 716, 2019.
- [25] H. Wei, R. Jafari, and N. Kehtarnavaz, "Fusion of video and inertial sensing for deep learning-based human action recognition," *Sensors*, vol. 19, no. 17, p. 3680, 2019.
- [26] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 3, pp. 226–239, 1998.
- [27] P. Emerson, "The original borda count and partial voting," *Social Choice and Welfare*, vol. 40, no. 2, pp. 353–358, 2013.
- [28] M. van Erp, L. Vuurpijl, and L. Schomaker, "An overview and comparison of voting methods for pattern recognition," in *Workshop on Frontiers in Handwriting Recognition*, 2002.
- [29] M. Ramanathan, J. Kochanowicz, and N. M. Thalmann, "Combining pose-invariant kinematic features and object context features for rgb-d action recognition," *International Journal of Machine Learning and Computing*, vol. 9, no. 1, pp. 44–50, 2019.
- [30] G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *International ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 758–759.
- [31] N. Damer, P. Terhöst, A. Braun, and A. Kuijper, "General borda count for multi-biometric retrieval," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 420–428.
- [32] R. Sharma, S. Das, and P. Joshi, "Rank level fusion in multibiometric systems," in *Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*. IEEE, 2015, pp. 1–4.
- [33] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi, "Head, eye, and hand patterns for driver activity recognition," in *International Conference on Pattern Recognition*, 2014, pp. 660–665.
- [34] A. Jain, H. S. Koppula, S. Soh, B. Raghavan, A. Singh, and A. Saxena, "Brain4cars: Car that knows before you do via sensory-fusion deep learning architecture," *arXiv preprint arXiv:1601.00740*, 2016.
- [35] C. Wang, H. Yang, and C. Meinel, "Exploring multimodal video representation for action recognition," in *International Joint Conference on Neural Networks*. IEEE, 2016, pp. 1924–1931.
- [36] A. Roitberg, T. Pollert, M. Haurilet, M. Martin, and R. Stiefelhausen, "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, pp. 0–0.
- [37] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5492–5501.
- [38] A. Roitberg, M. Haurilet, S. Reiß, and R. Stiefelhausen, "Cnn-based driver activity understanding: Shedding light on deep spatiotemporal representations," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–6.
- [39] J. F. Masakuna, S. W. Utete, and S. Kroon, "Performance-agnostic fusion of probabilistic classifier outputs," in *International Conference on Information Fusion (FUSION)*. IEEE, 2020, pp. 1–8.
- [40] Y. Wang, J. Li, and F. Metzke, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 31–35.
- [41] P. Drotár, M. Gazda, and J. Gazda, "Heterogeneous ensemble feature selection based on weighted borda count," in *2017 9th International Conference on Information Technology and Electrical Engineering (ICITEE)*. IEEE, 2017, pp. 1–4.