Sensor Fusion and Multimodal Learning for Robotic Grasp Verification

Priteshkumar Gohil[1], Santosh Thoduka[1] and Paul G. Plöger[1]

Abstract— **Robots with different sensors could help in understanding different aspects of the environment they are working in; however, each sensor modality is often processed individually and information from other sensors is not utilized jointly. One of the reasons is different sampling rates and dimensions of different modalities. In this paper, we use multimodal data fusion techniques such as early, late and intermediate fusion for grasp failure identification using four different 3D convolution-based multimodal neural networks (3D-MNN). Our results on a visual-tactile dataset shows that the performance of the classification task is improved while using multimodal data. In addition, a neural network trained with 30:22 train-test split of multimodal data achieved accuracy comparable to a network trained with 78:22 train-test split of unimodal data.**

I. INTRODUCTION

Machine learning and deep learning models are widely used for complex tasks in robotics such as manipulation, navigation, object detection and classification. Vision is the dominant modality used in several learning-based applications. Although vision-based applications have achieved state-of-the-art results, they typically process and focus on improving the performance of a task using a single modality. Moreover, relying only on a single visual modality, such as an RGB camera, may impose multiple challenging situations such as poor illumination, motion or lens blur and limited field of view.

To address these problems, researchers proposed a human-like approach to learn a task from multiple sensory modalities; also called multimodal learning [1]–[4]. For example, multiple sensors can capture the occurrence of an event in different ways. In addition, multimodal perception allows models to depict multiple sources of information to learn a task which is more robust than unimodal data [5]. Such models can also boost accuracy by 10% to 25%, can adapt to different environments and work in the absence of other input modalities. For example, a model trained on multiple modalities can produce good outputs even if one of the modalities is missing.

However, teaching models to process and combine multimodal data is challenging because of multiple reasons such as different dimensions of sensor data, heterogeneity gap between the original data and extracted features, effective sensor fusion and limited number of multimodal datasets [5]. Therefore, in this paper, we propose a 3D convolution-based multimodal neural network (3D-MNN) with different fusion techniques for the grasp failure detection task. We use visual, tactile and finger joint positions from a multimodal visual-tactile dataset [6]. We further demonstrate the effectiveness of the proposed network by performing comparative experiments of the 3D-MNN model with different sensor inputs and different fusion strategies.

II. RELATED WORK

Combining multiple modalities at different levels such as data level, feature level or decision level is called data or sensor fusion [7]. Some of the techniques for data fusion are Kalman filter, maximum likelihood, Bayesian networks, and neural networks. Bayesian networks, Kalman filter and maximum likelihood techniques are used for state estimation or tracking the target under moving conditions [8], [9]. Neural network based fusion approaches have been more popular recently because of their speed and accuracy.

Sensor fusion is further divided into four categories as seen in Figure 1. In early fusion, data are combined at the early stage or sometimes before extracting features. However, temporal and context dependencies or interaction between modalities is lost in early fusion. Intermediate fusion combines the learned features of different modalities which helps in preserving temporal and contextual information for individual modalities. There exists numerous approaches for feature level fusion such as concatenation [1], [2], [10], tensor product [11], temporal fusion using Long short-term memory (LSTM) [12] and element-wise summation [13]. Multilayer fusion methods combine sensor data at multiple layers of feature extraction. Studies such as [14], [15] perform multilayer fusion using gated units. In late fusion, each input modality is trained using a separate model and the decision of each model is combined in different ways; for example using voting, maximum, minimum, sum, product, median or multiplicative fusion. It can be seen as an ensemble of classifiers.

In grasp verification, Calandra et al. [16] shows that using vision and tactile modalities together improves the grasp success rate. Lee et al. [1] obtain the joint representation of video, force, velocity and position of a robotic arm to learn manipulation control policies effectively. Inceoglu et al. [3] identify grasp failure from audio, RGB and depth image modalities using a 2D Convolutional Neural Network (CNN). However, these studies use multimodal information at a specific time instance and do not utilize temporal information of input modalities. Cui et al. [2] uses a 3D CNN to assess the grasp state by fusing visual and tactile data. Their approach is similar to ours; however, they perform only intermediate fusion and do not provide a comparison with other fusion techniques. Nevertheless, to best of our knowledge, none

[1] Autonomous Systems, Department of Computer Science, Hochschule Bonn-Rhein-Sieg, Germany pritesh.gohil@inf.h-brs.de, santosh.thoduka@h-brs.de, paul.ploeger@h-brs.de

of the previous studies provide a comparative evaluation of different fusion strategies on grasp verification task using multimodal data. Hence, we present 3D-MNN to understand the effectiveness of different fusion strategies.

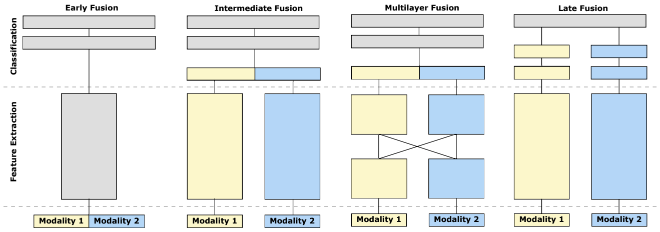


Fig. 1: Data fusion strategies explained with two modalities using a high-level diagram. Gray color indicates the level where different modalities interact with each other.

III. METHODOLOGY

Our goal is to predict the current grasp state as success or failure from visual, tactile and joint position data using different fusion methods. Consider visual $\{X_{v1}, X_{v2}, \cdots, X_{vm}\}$, tactile $\{X_{t1}, X_{t2}, \cdots, X_{tm}\}$ and position $\{X_{p1}, X_{p2}, \cdots, X_{pm}\}$ input sequences. We first extract the features (F_{vtp}) for each modality using visual (M_v), tactile (M_t), and position (M_p) neural networks. Then extracted features are passed through binary classification function $\mathbb{F}$ to predict the grasp state Y. The problem is formulated in Equations (1) and (2) as in [2].

$$\begin{aligned} F_{vtp} = M_v(X_{v1}, X_{v2}, \cdots, X_{vm}) \bigoplus \\ M_t(X_{t1}, X_{t2}, \cdots, X_{tm}) \bigoplus \\ M_p(X_{p1}, X_{p2}, \cdots, X_{pm}) \end{aligned} \tag{1}$$

$$Y = \mathbb{F}(F_{vtp}), Y \in 0, 1 \tag{2}$$

Where, m is the number of frames in each modality and output is 0 for grasp failure or 1 for success. This problem is further decomposed in early (3D-MNN$_{early}$), intermediate (3D-MNN$_{int}$) and late fusion (3D-MNN$_{late}$) with minor changes in the network structure. M_v, M_t, M_p are implemented as 3D-CNNs and the classification function $\mathbb{F}$ is implemented by Fully-Connected (FC) layers.

A. Dataset

We use the visual-tactile multimodal dataset [6] where the robot is performing pickup task for 10 different objects. The robotic hand is equipped with 16 tactile sensors and 8 motors to control fingers. RGB videos are recorded using third person front and left cameras. We select RGB front video, tactile data and joint position data containing 1658 grasp experiments of approximately 18 seconds each. Figure 2 shows an example where a coffee tin slips through the hand during a grasp.

B. Model Details

An overview of the proposed models is shown in Figure 3 and the detailed architecture of one of the models, 3D-MNN$_{\text{int}}$, is shown in Figure 4. Most of the network components such as 3D-CNN and FC layers are reused between different models; except for the change in network input modality dimension. In all our experiments, the frame length is 18 for each modality; which corresponds to 1 second of the execution. Moreover, all three input modalities are time-aligned with each other.

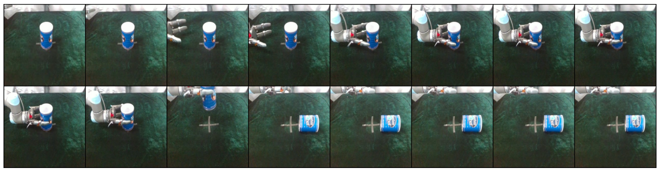


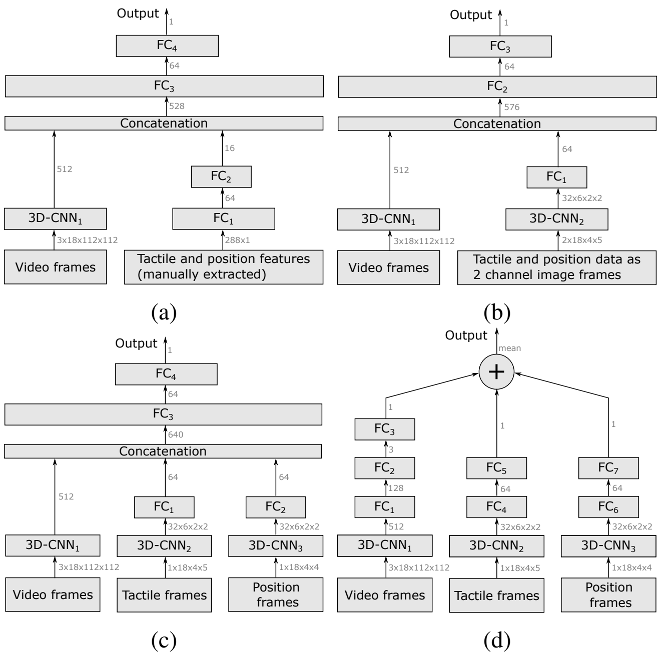
Fig. 2: Visualization of video frames from the front camera during a pickup task. Frames are visualized at every second starting from the top left and ending at the bottom right [6].

In 3D-MNN$_{\text{early1}}$ (see Figure 3a), we aim to fuse tactile and position sensor data in the early stage by concatenating hand extracted features such as summation, minimum, maximum, mean, standard deviation, kurtosis, skewness, principal component analysis (PCA), 20th, 40th, 60th, 80th percentile from each sensor channel. These statistical features are most commonly used and describes the time series data with few numbers. Finally, we have 18 image frames of shape $3 \times 112 \times 112$, 192×1 vector for tactile and 96×1 vector for the position data input to the model. The features from the visual modality of shape 512×1 are concatenated with temporal features of shape 16×1. The final grasp state output is classified by two FC layers.

For the other three models, we use a visual representation of the tactile sensor and joint positions, because of matrix-like sensor placement. The tactile sensors on the robotic hand are spatially correlated to each other, hence converting the tactile channels into a 2D grid like structure allows us to use CNNs to extract spatial features as well. The same applies to the joint position data. In 3D-MNN$_{\text{early2}}$ (see Figure 3b), a tactile and position image of shape 4×5 is combined in the early stage prior to feature extraction. Visual features of shape 512×1 and temporal features of shape 64×1 are concatenated before giving input to the classification head. In 3D-MNN$_{\text{int}}$ (see Figure 3c), features from all three input modalities are extracted separately and spatio-temporal features are fused in the intermediate stage. Finally, in the late fusion model 3D-MNN$_{\text{late}}$ (see Figure 3d), feature extraction and classification for each input modality is done separately, and the final decision from each modality is fused by calculating the mean value.

IV. EXPERIMENTAL RESULTS

The proposed models for robotic grasp verification were evaluated on the visual-tactile dataset [6]. We split the available data into train and test sets in three ways: i) TTrandom: data are randomly split with a 80:20 train-test ratio; ii) TTobject: data are split based on the object being manipulated to test generalizability to unseen objects (78:22 ratio) and iii) TTsides: data are split based on the grasp strategy (top and right grasp for training, and back grasp



(a) (b)

(c) (d)

Fig. 3: Proposed multimodal data fusion frameworks. (a) and (b) are 3D-MNN$_{\text{early1}}$ and 3D-MNN$_{\text{early2}}$ network respectively with two different early fusion approaches. (c) 3D-MNN$_{\text{int}}$ network with intermediate fusion. (d) 3D-MNN$_{\text{late}}$ network with late fusion.

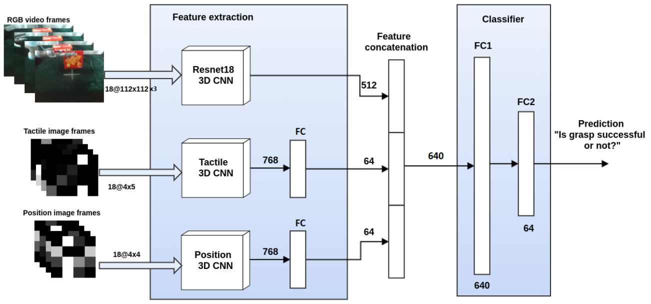


Fig. 4: 3D-MNN$_{\text{int}}$ architecture using 3D ResNet-18 as a feature extractor for visual and two layer 3D-CNN for tactile and joint position modalities. The input to the network is 18 frames from all the modalities, where visual is 3 channel (RGB) and tactile or joint position data are single channel grayscale images.

for testing) with a 68:22 ratio. We evaluate our approach using accuracy, precision, recall and F1-score.

A. Comparison Between Different Fusion Approaches

Table I summarizes results of different fusion approaches using TTobject data split. The results are worse when using only visual modality, because the decision of the classifier was biased to features of visual objects. The accuracy and precision has slightly improved for 3D-MNN$_{\text{early1}}$ network compared to unimodal network. Whereas, 3D-MNN$_{\text{early2}}$ and 3D-MNN$_{\text{late}}$ do not show much improvement when all modalities are used. 3D-MNN$_{\text{int}}$ network achieves the best performance after reducing the dimension of the feature map of the visual modality. Then it is concatenated with other modalities. This helped us to improve accuracy from 86.9% to 90.8%. The overall comparison reveals that intermediate fusion performs better than early or late fusion. Therefore, we use 3D-MNN$_{\text{int}}$ network for all other experiments to evaluate other properties of the network.

TABLE I: Comparison between individual modalities and proposed fusion approaches using TTobject split.

Approach	Modality			Accuracy	Precision	Recall	F1-score
	Video	Tactile	Pos				
3D-CNN (No Fusion)	✓			0.801	0.801	0.999	0.889
		✓		0.893	0.892	0.986	0.937
			✓	0.860	0.855	0.994	0.919
3D-MNN$_{\text{early1}}$		✓	✓	0.858	0.860	0.981	0.917
	✓	✓	✓	0.863	0.863	0.985	0.920
3D-MNN$_{\text{early2}}$		✓	✓	0.878	0.875	0.989	0.929
	✓	✓	✓	0.868	0.863	0.993	0.923
3D-MNN$_{\text{int}}$	✓	✓		0.865	0.862	0.991	0.922
		✓	✓	0.894	0.892	0.986	0.937
	✓	✓	✓	0.869	0.864	0.993	0.924
	✓†	✓†	✓†	**0.908**	**0.905**	**0.988**	**0.945**
3D-MNN$_{\text{late}}$	✓	✓		0.856	0.851	0.994	0.917
		✓	✓	0.890	0.902	0.967	0.933
	✓	✓	✓	0.865	0.894	0.944	0.918

† Reducing the dimension of visual feature map from 512×1 to 16×1 before the feature fusion.

B. Training Data are Reduced to 30% in TTobject Split

This experiment is interesting especially to observe the network performance when we do not have a large dataset. Therefore, training data are reduced to only 30 % from 78% by removing samples randomly from the training set and while the test set remains unchanged. As seen in Table II, the performance for most modality variants decreases as expected compared to the Table I (especially results from the tactile modality). Video+tactile+joint position still performs the best, and tactile+joint position in fact shows an improvement in performance compared to Table I. Hence, this experiment shows that learning with multimodal data can be maximized even with a small dataset.

TABLE II: Comparison of our results using TTobject split where training set is reduced to 30% from 78%.

Classifier	Modality			Accuracy	Precision	Recall	F1-score
	Video	Tactile	Pos				
3D-CNN (No Fusion)	✓			0.801	0.801	**0.999**	0.889
		✓		0.849	0.847	0.991	0.913
			✓	0.856	0.851	0.994	0.917
3D-MNN$_{\text{int}}$	✓	✓		0.874	0.869	0.991	0.926
		✓	✓	0.889	0.889	0.983	0.934
	✓	✓	✓	**0.891**	**0.889**	0.987	**0.935**

C. Training and Testing using Data Split TTrandom

This experiment helps to compare our result with the classical machine learning techniques such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes (NB) and LSTM as reported in [6]. They use only tactile data for their experiments and report only accuracy. Experiments using TTobject split perform worse than the random split, however it outperforms LSTM or SVM. Table III shows that the 3D-MNN$_{\text{int}}$ network using random split outperforms classical machine learning techniques. Only the video modality achieves the same accuracy as LSTM. Results are improved as more modalities are used. Finally, using all three modalities (video+tactile+joint position) provides the best performance with the TTrandom split.

TABLE III: Comparison of our results using TTrandom split. Our approach with only tactile data outperforms the results from SVM classifier.

Classifier	**Modality**			**Accuracy**	**Precision**	**Recall**	**F1-score**
	Video	**Tactile**	**Pos**				
KNN [6]		✓		0.797	-	-	-
SVM [6]		✓		0.847	-	-	-
NB [6]		✓		0.688	-	-	-
LSTM [6]		✓		0.846	-	-	-
3D-CNN (No Fusion)	✓			0.846	0.853	**0.980**	0.912
		✓		0.892	0.919	0.951	0.935
			✓	0.916	0.937	0.962	0.949
3D-MNN$_{\text{int}}$	✓	✓		0.926	0.945	0.966	0.955
	✓†	✓†	✓†	0.951	0.979	0.962	0.970
	✓	✓	✓	**0.957**	**0.981**	0.966	**0.973**

† Network is trained using 50% of training data

V. CONCLUSIONS

A network 3D-MNN is proposed in this paper to evaluate the robotic object grasp state by employing different fusion methods on visual, tactile and joint position modalities. On a visual-tactile dataset, we compare results from early, intermediate and late fusion methods; and observed that intermediate fusion with all three modalities has the highest F1-score. Also, the multimodal network's performance using 30:22 train-test split is as good as results obtained by 78:22 train-test split. Hence, the proposed network can learn even with less data and reduces the problem of over-fitting.

Future work can explore the use of attention-based networks [17], [18] for learning to attend to the important modalities. Further experiments can be conducted to test the limit of the network in terms of the number of modalities it can fuse and exploring learning error minimization tricks for multimodal networks.

ACKNOWLEDGMENT

We gratefully acknowledge the support of the Bonn-Aachen International Center for Information Technology.

REFERENCES

[1] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8943–8950.

[2] S. Cui, R. Wang, J. Wei, F. Li, and S. Wang, "Grasp state assessment of deformable objects using visual-tactile fusion perception," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 538–544.

[3] A. Inceoglu, E. E. Aksoy, A. C. Ak, and S. Sariel, "Fino-net: A deep multimodal sensor fusion framework for manipulation failure detection," *CoRR*, vol. abs/2011.05817, 2020.

[4] D. Jiang, G. Li, Y. Sun, J. Hu, J. Yun, and Y. Liu, "Manipulator grabbing position detection with information fusion of color image and depth image using deep learning," *Journal of Ambient Intelligence and Humanized Computing*, Jan 2021.

[5] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63 373–63 394, 2019.

[6] T. Wang, C. Yang, F. Kirchner, P. Du, F. Sun, and B. Fang, "Multimodal grasp data set: A novel visualtactile data set for robotic manipulation," *International Journal of Advanced Robotic Systems*, vol. 16, no. 1, p. 1729881418821571, 2019.

[7] D. Lahat, T. Adalı, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, Aug 2015.

[8] V. Kubelka, L. Oswald, F. Pomerleau, F. Colas, T. Svoboda, and M. Reinstein, "Robust data fusion of multimodal sensory information for mobile robots," *Journal of Field Robotics*, vol. 32, no. 4, pp. 447–473, 2015.

[9] A. Singhal and C. R. Brown, "Dynamic Bayes net approach to multimodal sensor fusion," in *Sensor Fusion and Decentralized Control in Autonomous Robotic Systems*, P. S. Schenker and G. T. McKee, Eds., vol. 3209, International Society for Optics and Photonics. SPIE, 1997, pp. 2 – 10.

[10] T. Rotondo., G. M. Farinella., V. Tomaselli., and S. Battiato, "Action anticipation from multimodal data," in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 4, INSTICC. SciTePress, 2019, pp. 154–161.

[11] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2247–2256.

[12] A. Gandhi, A. Sharma, A. Biswas, and O. Deshmukh, "Gethr-net: A generalized temporally hybrid recurrent neural network for multimodal information fusion," in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016, pp. 883–899.

[13] J.-H. Choi and J.-S. Lee, "Embracenet: A robust deep learning architecture for multimodal classification," *Information Fusion*, vol. 51, pp. 259–270, 2019.

[14] J. Arevalo, T. Solorio, M. M. y Gmez, and F. A. Gonzlez, "Gated multimodal units for information fusion," 2017.

[15] J. Bimbo, S. Luo, K. Althoefer, and H. Liu, "In-hand object pose estimation using covariance-based tactile to geometry matching," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 570–577, 2016.

[16] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?" *CoRR*, vol. abs/1710.05512, 2017.

[17] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo, "Deep multimodal representation learning from temporal data," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5066–5074.

[18] M. M. Islam and T. Iqbal, "Multi-gat: A graphical attention-based hierarchical multimodal representation learning approach for human activity recognition," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1729–1736, 2021.