Hand Pose Estimation using Convolution Neural Networks

Ms. Pooja Dewangan, Mrs. Aditi Mishra, Mr. L P Bhaiya
(Department of Computer Science & Eng, Bharti College of Eng. & Tech. Chhattisgarh, India.)

Abstract: **Touchless hand gesture recognition systems have become vital in automotive user interfaces as they improve safety and luxury. numerous laptop vision algorithms have utilized color and depth cameras for hand gesture recognition, however sturdy classification of gestures from totally different subjects performed below wide varied lighting conditions continues to be difficult. we have a tendency to propose AN algorithmic rule for drivers' hand gesture recognition from difficult depth and intensity knowledge mistreatment 3D convolution neural networks. Our resolution combines data from multiple spatial scales for the ultimate prediction. It conjointly employs spatiotemporal knowledge augmentation for more practical coaching and to scale back potential overfitting. Our technique achieves an accurate classification rate of seventy seven.5% on the exam challenge dataset.**

Index Terms – **Hand Gesture Recognition, SLR, CNN, ML.**

I. Introduction

Hand gesture recognition is vital for coming up with touchless interfaces in cars. Such interfaces enable drivers to target driving whereas interacting with different controls, e.g., audio and air con, and therefore improve drivers' safety and luxury. within the last decade, several vision based dynamic hand gesture recognition algorithms were introduced [11, 16]. to acknowledge gestures, totally different options like handcrafted spatiotemporal descriptors [23] and articulated models [9], were used. As gesture classifiers, hidden Andrei Markov models [20], conditional random fields [24] and support vector machines (SVM) [4] are wide used. However, sturdy classification of gestures below wide varied lighting conditions, and from totally different subjects continues to be a difficult downside [25, 1, 15].

To improve classification accuracy, gesture recognition ways with multimodal sensors were introduced [14, 21, 12, 5, 13]. Neverova et al. with success combined RGBD knowledge from the hand region with upper body skeletal motion knowledge mistreatment convolution neural networks (CNNs) for recognizing twenty Italian signing gestures [13]. However, their technique was supposed for gestures performed inside solely. OhnBar and Trivedi evaluated numerous hand Crafted spatiotemporal options and classifiers for incar handgesture recognition with RGBD knowledge [14]. They reported the most effective performance with a mix of bar graph of gradient (HOG) options ANd an SVM classifier. Molchanov et al. [12] amalgamated data of hand gestures from depth, color and radiolocation sensors and together trained a convolution neural network with it. They incontestable fortunate classification results for wide varied lighting conditions, that actuated our work.

Recently, classification with deep convolution neural networks has been fortunate in numerous recognition challenges [2, 8, 10, 18]. Multicolumn deep CNNs that use multiple parallel networks are shown to enhance recognition rates of single networks by 30.80% for numerous image classification tasks [3]. Similarly, for giant scale video classification, Karpathy et al. [7] discovered the most effective results on combining CNNs trained with 2 separate streams of the initial and spatially cropped video frames.

Several authors have stressed the importance of mistreatment several various coaching examples for CNNs [8, 17, 19]. they need projected knowledge augmentation methods to forestall CNNs from overfitting once coaching with datasets containig restricted diversity. Krizhevsky et al. [8] utilized translation, horizontal flipping and RGB changeful of the coaching and testing pictures for classifying them into a thousand classes. Simonyan and Zisserman [19] utilized similar spatial augmentation on every video frame to coach CNNs for videobased act recognition. However, these knowledge augmentation ways were restricted to spatial variations solely. to feature variations to video sequences containing dynamic motion, Pigou et al. [17] temporally translated the video frames additionally to applying spatial transformations.

In this paper, we have a tendency to introduce a hand gesture recognition system that utilizes depth and intensity channels with 3D convolution neural networks. actuated by Molchanov et al. [12], we have a tendency to interleave the 2 channels to create normalized spatiotemporal volumes, and train 2 separate sub networks with these volumes. to scale back potential overfitting and improve generalization of the gesture classifier, we have a tendency to propose an efficient spatiotemporal knowledge augmentation technique to deform the input volumes of hand gestures. The augmentation technique conjointly incorporates existing spatial augmentation techniques [8]. This work bears similarities to the multisensory approach of Molchanov et al. [12], however differs within the utilization of 2 separate sub networks and knowledge augmentation.

We demonstrate that our system, with 2 sub networks, that employs spatiotemporal knowledge augmentation for coaching, outperforms each one CNN and also the baseline feature based algorithmic rule [14] on the exam challenge's dataset.

II. METHODOLOGY

We use a convolution neural network classifier for dynamic hand gesture recognition. Sec. 2.1, shortly describes the exam challenge's hand gesture dataset employed in this paper. Sec. 2.2 to 2.4 describe the preprocessing steps required for our model, the main points of the classifier and also the coaching pipeline for the 2 sub networks (Fig. 1). Finally, we have a tendency to introduce a spatiotemporal knowledge augmentation technique in Sec. 2.5, and show however it's combined with spatial transformations.

2.1. Dataset

The exam challenge was organized to judge and advance the state of the art in multimodal dynamic hand gesture recognition below difficult conditions (with variable lighting and multiple subjects). The exam challenge's dataset contains eight5 intensity and depth video sequences of nineteen totally different dynamic hand gestures performed by 8 subjects within a vehicle [14]. each channels were recorded with the Microsoft Kinect device and have a resolution of one hundred fifteen 250 pixels. The dataset was collected below varied illumination conditions. The gestures were performed either with the correct hand by subjects within the driver's seat or with the left by subjects within the front passenger's seat. The hand gestures involve hand and/or finger motion.

2.2. Preprocessing

Each hand gesture sequence in the VIVA dataset has a different duration. To normalize the temporal lengths of the gestures, we first resampled each gesture sequence to 32 frames using nearest neighbor interpolation (NNI) by dropping or repeating frames [12]. We also spatially down sampled the original intensity and depth images by a factor of 2 to 57 125 pixels. We computed gradients from the intensity channel using the Sobel operator of size 3X3 pixels. Gradients helped to improve robustness to the different illumination conditions present in the dataset. We normalized each channel of a particular gesture's video sequence to be of zero mean and unit variance. This helped our gesture classifier converge faster. The final inputs to the gesture classifier were 57 125 32 sized columns containing interleaved image gradient and depth frames.

2.3. Classifier

Our convolution neural network classifier consisted of two sub networks (Fig. 1): a high resolution network (HRN) and low resolution network (LRN), with network parameters WH and WL, respectively. Each network, with parameters W, produced class membership probabilities (P (Cjx; W) for classes C given the gesture's observation x. We multipled the class membership probabilities from the two networks element wise to compute the final class membership probabilities for the gesture classifier:

$$P\,(Cjx) = P\,(Cjx;\ WL)\ \ P\,(Cjx;\ WH\,): \qquad (1)$$

We predicted the class label c = arg max P (Cjx). The networks contained more than 1:2 million trainable parameters.

The high resolution network consisted of four 3D convolution layers, each of which was followed by the maxpooling operator. Fig. 1 shows the sizes of the convolution kernels, volumes at each layer, and the pooling operators. We input the output of the fourth 3D convolution layer to two fully connected layers (FCLs) with 512 and 256 neurons, respectively. The output of this high resolution network was a softmax layer, which produced class membership probabilities (P (Cjx; WH)) for the 19 gesture classes.

We input a spatially down sampled (via NNI) gesture volume of 28 62 32 interleaved depth and image gradient values to the low resolution network. Similar to the HRN, LRN also comprised of a number of 3D convolution layers, each followed by a maxpooling layer, two FCLs, and an output softmax layer that estimated the class membership probability P (Cjx; WL) values (Fig. 1).

All the layers in the networks, except for the softmax layers, had rectified linear unit (ReLU) activation functions:

$$f(z) = \max(0;\ z): \qquad (2)$$

2.4. Training

The process of training a CNN involves the optimization of the network's parameters W to minimize a cost function for the dataset D. We selected negative log likelyhood as the cost function:

$$\mathcal{L}(\mathcal{W},\mathcal{D}) = -\frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|}\log\left(P(C^{(i)}|x^{(i)},\mathcal{W})\right). \qquad (3)$$

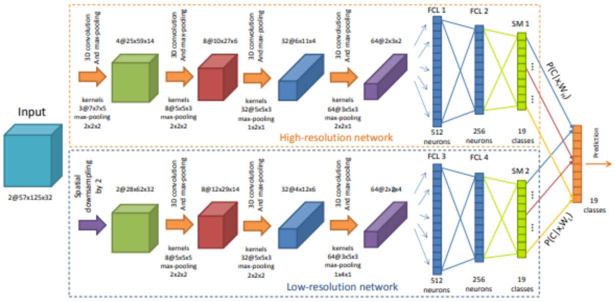

Input
2@57x125x32
3D convolution And max-pooling
4@25x59x14
kernels 3@7x7x5 max-pooling 2x2x2
8@10x27x6
kernels 8@5x5x3 max-pooling 2x2x2
32@6x11x4
kernels 32@5x5x3 max-pooling 1x2x1
64@2x3x2
kernels 64@3x5x3 max-pooling 2x2x1
High-resolution network
FCL 1
FCL 2
SM 1
512 neurons
256 neurons
19 classes
P(C|x,W_H)
Prediction
19 classes
Spatial downsampling by 2
2@28x62x32
8@12x29x14
kernels 8@5x5x5 max-pooling 2x2x2
32@4x12x6
kernels 32@5x5x3 max-pooling 2x2x2
64@2x2x4
kernels 64@3x5x3 max-pooling 1x4x1
Low-resolution network
FCL 3
FCL 4
SM 2
P(C|x,W_L)

Figure 1: Overview of our CNN classifier

We performed optimization via stochastic gradient descent with mini batches of 40 and 20 training samples for the LRN and the HRN, respectively.

III. RESULT AND DISCUSSION

We evaluated the performance of our dynamic hand gesture recognition system using leave-one-subject-out cross validation on the VIVA challenge's dataset [14]. We used data from one of the 8 subjects for testing and trained the classifier with data from the 7 remaining subjects; we repeated this process for each of the 8 subjects and averaged the accuracy. Fig. 6 shows the performance of the LRN during training. We applied various forms of regularization to the network in order to prevent overfitting even after a large number of training epochs. Data augmentation and drop-out were key components to successful generalization of the classifier.

The correct classification rates for our gesture recognition system are listed in Table 1. We compared our classifier to the baseline method proposed by Ohn-Bar and Trivedi [14], which employs HOG+HOG2 features. Both the low and high resolution convolution neural networks that we proposed, outperformed Ohn-Bar and Trivedi's method by 9.8% and 5.5%, respectively. Furthermore, the final classifier that combined the outputs of LRN and HRN outperformed the baseline method by 13.0%. Moreover, 52% of the final classifier's errors were associated with the second most probable class. The results indicate that our CNN based classifier for in-car dynamic hand gesture recognition considerably outperforms approaches that employ handcrafted features.

Method	HOG	LRN	HRN	LRN+HRN	NEW
Mean	64.5%	74.4%	70.0%	77.5%	93.2%
Std	16.9%	8.9%	7.8%	7.9%	7.9%

Table 1: The classification results for leave one subject out cross validation Both the LRN and the HRN outperformed the HOG based approach [14]. Our final classifier, which combined the LRN and the HRN resulted in the best performance.

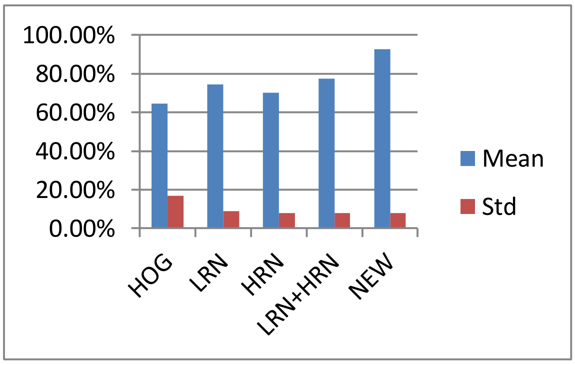


Figure 1: Performance chart according to result.

CNNs have been shown to be effective at combining data from different sensors [12]. We compared the performance of early and late fusion of sensors in the CNN architecture. In Table 2, we present the correct classification rates for the LRN trained with different input modalities. Observe that, individually, the depth data (accuracy = 65%) performed better than the intensity data (accuracy = 57%). On element wise late multiplying the class membership probabilities of the two LRNs trained individually with the two modalities, the accuracy improved to 70:4%. However, the highest accuracy of 74:4% was obtained when the LRN was trained with interleaved depth and image gradient frames as input.

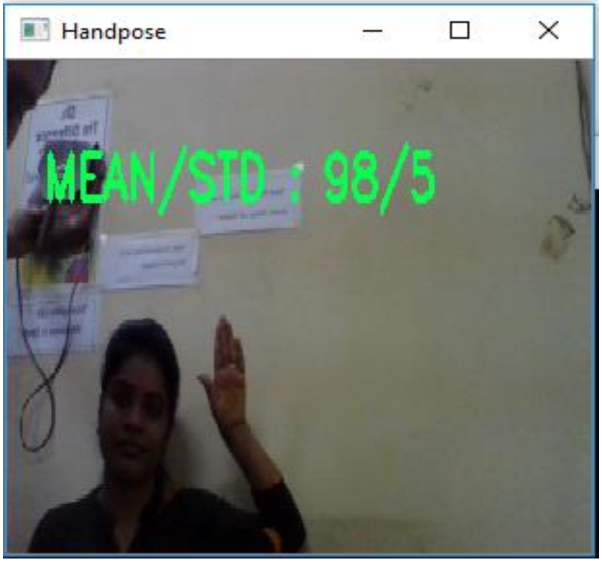


Figure 2: Acquiring inputs from user using RGB Camera.

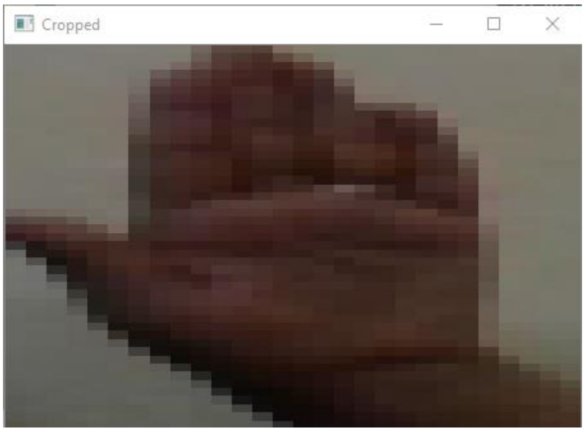


Figure 3: Cropping of hand from images.

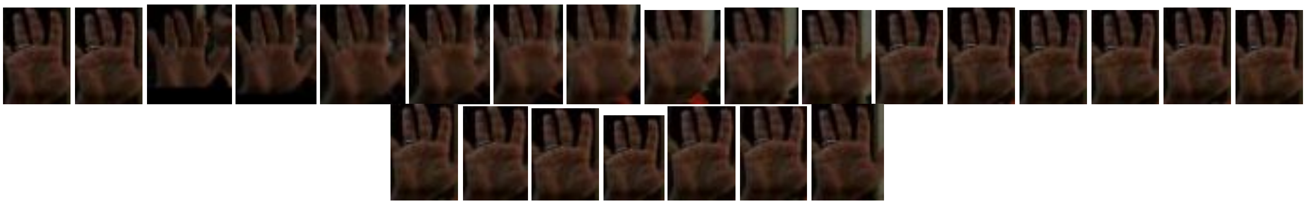
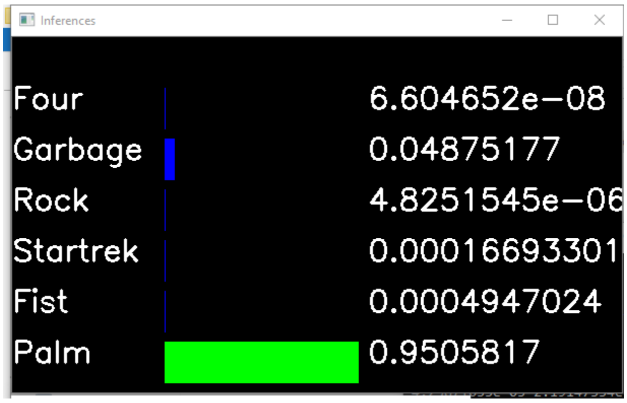
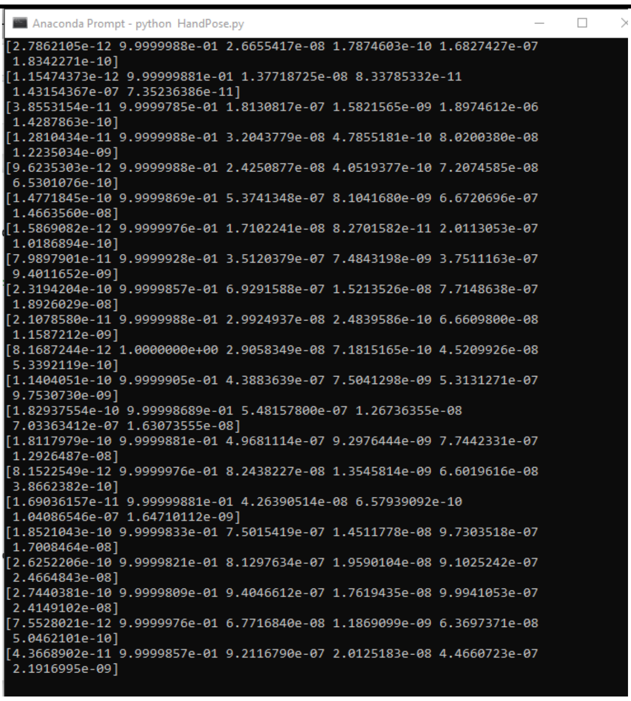


Figure 4: Palm datasets



```
Four       6.604652e-08
Garbage    0.04875177
Rock       4.8251545e-06
Startrek   0.00016693301
Fist       0.0004947024
Palm       0.9505817
```

Figure 4: Plot chart as per result.



```
[2.7862105e-12 9.9999988e-01 2.6655417e-08 1.7874603e-10 1.6827427e-07
 1.8342271e-10]
[1.15474373e-12 9.99999881e-01 1.37718725e-08 8.33785332e-11
 1.43154367e-07 7.35236386e-11]
[3.8553154e-11 9.9999785e-01 1.8130817e-07 1.5821565e-09 1.8974612e-06
 1.4287863e-10]
[1.2810434e-11 9.9999988e-01 3.2043779e-08 4.7855181e-10 8.0200380e-08
 1.2235034e-09]
[9.6235303e-12 9.9999988e-01 2.4250877e-08 4.0519377e-10 7.2074585e-08
 6.5301076e-10]
[1.4771845e-10 9.9999869e-01 5.3741348e-07 8.1041680e-09 6.6720696e-07
 1.4663560e-08]
[1.5869082e-12 9.9999976e-01 1.7102241e-08 8.2701582e-11 2.0113053e-07
 1.0186894e-10]
[7.9897901e-11 9.9999928e-01 3.5120379e-07 7.4843198e-09 3.7511163e-07
 9.4011652e-09]
[2.3194204e-10 9.9999857e-01 6.9291588e-07 1.5213526e-08 7.7148638e-07
 1.8926029e-08]
[2.1078580e-11 9.9999988e-01 2.9924937e-08 2.4839586e-10 6.6609800e-08
 1.1587212e-09]
[8.1687244e-12 1.0000000e+00 2.9058349e-08 7.1815165e-10 4.5209926e-08
 5.3392119e-10]
[1.1404051e-10 9.9999905e-01 4.3883639e-07 7.5041298e-09 5.3131271e-07
 9.7530730e-09]
[1.82937554e-10 9.99998689e-01 5.48157800e-07 1.26736355e-08
 7.03363412e-07 1.63073555e-08]
[1.8117979e-10 9.9999881e-01 4.9681114e-07 9.2976444e-09 7.7442331e-07
 1.2926487e-08]
[8.1522549e-12 9.9999976e-01 8.2438227e-08 1.3545814e-09 6.6019616e-08
 3.8662382e-10]
[1.69036157e-11 9.99999881e-01 4.26390514e-08 6.57939092e-10
 1.04086546e-07 1.64710112e-09]
[1.8521043e-10 9.9999833e-01 7.5015419e-07 1.4511778e-08 9.7303518e-07
 1.7008464e-08]
[2.6252206e-10 9.9999821e-01 8.1297634e-07 1.9590104e-08 9.1025242e-07
 2.4664843e-08]
[2.7440381e-10 9.9999809e-01 9.4046612e-07 1.7619435e-08 9.9941053e-07
 2.4149102e-08]
[7.5528021e-12 9.9999976e-01 6.7716840e-08 1.1869099e-09 6.3697371e-08
 5.0462101e-10]
[4.3668902e-11 9.9999857e-01 9.2116790e-07 2.0125183e-08 4.4660723e-07
 2.1916995e-09]
```

Figure 5: Fetching Result as per hand pose.

In Table 3, we present the correct classification rates for the low resolution network with different forms of data augmentation. We observed that the training error increased on enabling data augmentation. However, the test error decreased. This demonstrates that the proposed data augmentation method successfully reduced overfitting and improved generalization of the gesture classifier. Additionally, we observed that image gradients increased the final correct classification rate by 1:1%, and spatial and temporal elastic deformations applied to the training data increased it by 1:2% and 1:72%, respectively.

Table 4 shows the confusion matrix of our proposed final classifier. Our classifier often confused between the Swipe and Scroll gestures performed along the same direction. Many gestures were misclassified as the Swipe down gesture. The Rotate CW/CCW gestures were difficult for the classifier. The classifier also had difficulties with distinguishing between the Swipe + and the Swipe X gestures. The Tap3 gesture produced 38% of the misclassifications.

The classifier's less confident decisions can be rejected by setting an empirical threshold. This helps to increase the correct classification rate, but at the cost of a greater number of missed gestures. Fig. 7 demonstrates this tradeoff for our gesture classifier at various confidence threshold values.

IV. CONCLUSION

We developed an effective method for dynamic hand gesture recognition with 3D convolution neural networks The proposed classifier uses a fused motion volume of normalized depth and image gradient values, and utilizes spatiotemporal data augmentation to avoid overfitting. By means of extensive evaluation, we demonstrated that the combination of low and high resolution sub-networks improves classification accuracy considerably. We further demonstrated that the proposed data augmentation technique plays an important role in achieving superior performance. For the challenging VIVA dataset, our proposed system achieved a classification rate of 77.5%. Our future work will include more adaptive selection of the optimal hyper parameters of the CNNs, and investigating robust classifiers that can classify higher level dynamic gestures including activities and motion contexts.

V. REFERENCES

[1] F. Althoff, R. Lindl, L. Walchshausl, and S. Hoch. Robust multimodal handand head gesture recognition for controlling automotive infotainment systems. VDI BERICHTE, 1919:187, 2005. 1

[2] D. C. Cires¸an, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In International Joint Conference on Artificial Intelligence, pages 1237–1242, 2011. 1

[3] D. Ciresan, U. Meier, and J. Schmidhuber. Multicolumn deep neural networks for image classification. In CVPR, pages 3642–3649. IEEE, 2012. 1

[4] N. Dardas and N. D. Georganas. Realtime hand gesture detection and recognition using bagoffeatures and support vector machine techniques. IEEE Transactions on Instrumentation and Measurement, 60(11):3592–3607, 2011. 1

[5] S. Escalera, X. Bar, J. Gonzlez, M. A. Bautista, M. Madadi, Reyes, V. Ponce, H. J. Escalante, J. Shotton, and Guyon. Chalearn looking at people challenge 2014: Dataset and results. In ECCVW, 2014. 1

[6] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing coadaptation of feature detectors. arXiv, 2012. 3

[7] Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. FeiFei. Largescale video classification with convolutional neural networks. In CVPR, pages 1725–1732. IEEE, 2014. 1

[8] Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, pages 1097–1105. 2012. 1, 2, 4

[9] J. J. LaViola Jr. An introduction to 3D gestural interfaces. In SIGGRAPH Course, 2014. 1

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradientbased learning applied to document recognition. In Proceedings of the IEEE, pages 2278–2324, 1998. 1

[11] S. Mitra and T. Acharya. Gesture recognition: A survey. IEEE Systems, Man, and Cybernetics, 37:311–324, 2007. 1

[12] P. Molchanov, S. Gupta, K. Kim, and K. Pulli. Multisensor System for Driver's Handgesture Recognition. In AFGR, 2015. 1, 2, 5

[13] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. Multiscale deep learning for gesture detection and localization. In ECCVW, 2014. 1

[14] E. OhnBar and M. Trivedi. Hand gesture recognition in real time for automotive interfaces: A multimodal visionbased approach and evaluations. Trans. ITS, (99):1–10, 2014. 1, 2, 5, 7

[15] F. ParadaLoira, E. GonzalezAgulla, and J. AlbaCastro. Hand gestures to control infotainment equipment in cars. In IEEE Intelligent Vehicles Symposium, pages 1–6, 2014. 1

[16] I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for humancomputer interaction: A review. PAMI, 19:677–695, 1997. 1

[17] L. Pigou, S. Dieleman, P.J. Kindermans, and B. Schrauwen. Sign language recognition using convolutional neural networks. In ECCVW, 2014. 1

[18] P. Y. Simard, D. Steinkraus, and J. C. Platt. J.c.: Best practices for convolutional neural networks applied to visual document analysis. In In: Intl Conference on Document Analysis and Recognition, pages 958–963, 2003. 1, 4

[19] K. Simonyan and A. Zisserman. Twostream convolutional networks for action recognition in videos. In NIPS, pages 568–576, 2014. 1, 4

[20] T. Starner, A. Pentland, and J. Weaver. Realtime american sign language recognition using desk and wearable computer based video. PAMI, 20(12):1371–1375, 1998. 1

[21] J. Suarez and R. R. Murphy. Hand gesture recognition with depth images: A review. In ROMAN, pages 411–417, 2012. 1

[22] Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In Proceedings of the 30th International Conference on Machine Learning (ICML13), pages 1139–1147, 2013. 3

[23] P. Trindade, J. Lobo, and J. Barreto. Hand gesture recognition using color and depth images enhanced with hand angular pose data. In Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE Conference on, pages 71–76, 2012. 1

[24] S. B. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In CVPR, pages 1521–1527, 2006. 1

[25] M. Zobl, R. Nieschulz, M. Geiger, M. Lang, and G. Rigoll. Gesture components for natural interaction with incar devices. In GestureBased Communication in HumanComputer Interaction, pages 448–459. Springer, 2004. 1