A Vision-based Transfer Learning Approach for Recognizing Behavioral Symptoms in People with Dementia

Zachary Wharton, Erik Thomas, Bappaditya Debnath and Ardhendu Behera
Department of Computer Science, Edge Hill University, Ormskirk, Lancashire, L39 4QP, UK
{zachary.wharton, erik.thomas}@go.edgehill.ac.uk, {debnathb,beheraa}@edgehill.ac.uk

Abstract

With an aging population that continues to grow, dementia is a major global health concern. It is a syndrome in which there is a deterioration in memory, thinking, behavior and the ability to perform activities of daily living. Depression and aggressive behavior are the most upsetting and challenging symptoms of dementia. Automatic recognition of these behaviors would not only be useful to alert family members and caregivers, but also helpful in planning and managing daily activities of people with dementia (PwD). In this work, we propose a vision-based approach that unifies transfer learning and deep convolutional neural network (CNN) for the effective recognition of behavioral symptoms. We also compare the performance of state-of-the-art CNN features with the hand-crafted HOG-feature, as well as their combination using a basic linear SVM. The proposed method is evaluated on a newly created dataset, which is based on the dementia storyline in ITVs Emmerdale episodes. The Alzheimer's Society has described it as a "realistic portrayal"[1] *of the condition to raise awareness of the issues surrounding dementia.*

1. Introduction

According to the Alzheimer's Society, around 46.8M people are living with dementia and the numbers will rise to 115.4M in 2050. One in three PwD shows aggressive behavior [2], which is very stressful and upsetting for the person with dementia and their carers. Depression is also common at all stages of dementia. It occurs in about 20–40% of PwD [1]. Identifying depression in PwD can be difficult. To date, there is no single test or questionnaire to detect the depression due to the complexities and multi-faceted nature of the condition.

The common approach to monitor and manage the above-mentioned behavioral symptoms is via direct observation by caregivers, family members and health care professionals. However, this is labor-intensive, subjective, time consuming, costly and could increase the workload of caregivers and health care professionals [10]. Recently, ambient technologies have been explored extensively in a variety of settings, such as "smart homes" and hospitals, for health monitoring. These technologies could be adapted into the early detection of behavioral symptoms that would aid caregivers and guide the headway of tailored interventions [39].

Most of the above-mentioned technologies often use on-body bio-sensing devices (e.g. actigraphs, accelerometers, biomarkers and biopatches) for measuring signals linking behavioral symptoms [5, 19, 7, 36]. However, it is suggested that PwD requires monitoring systems that are "unobtrusive, and preferably collected in a transparent way without patient intervention due to their cognitive impairment" [28]. Therefore, more recently researchers have explored monitoring systems involving unobtrusive methods that includes video surveillance using cameras and Kinect sensors [14, 21, 7]. Monitoring and recognition of aggression and depression using such systems is still very much in its infancy. This could be due to the challenge faced by the researchers to develop standard algorithms that can adequately and concisely recognize behavioral symptoms.

In this paper, we propose a novel method for recognizing behavioral symptoms involving aggression and depression. The proposed approach benefits from the power of *transfer learning* (TL) by using appearance features as deep CNN features, which are extracted from various state-of-the-art deep models (e.g. VGG16 [32], Inception-V3 [35] and Inception ResNet-V2 [34]). We also explore the various level of abstraction by exploring different extraction points in a given CNN model (e.g. VGG16 [32]). This work includes the following novel contributions:

- To our knowledge, we are the first to report vision-based recognition of behavioral symptoms (*aggressive*, *depressive*, *happy* and *neutral*) in PwD.
- We demonstrate the effectiveness of TL using different state-of-the-art deep CNN models for recognizing behavioral symptoms in PwD. We evaluate various com-

[1] www.bbc.co.uk/news/entertainment-arts-38389953

978-1-5386-9294-3/18/$31.00 ©2018 IEEE

binations of deep CNN features using SVM.

- We introduce a novel image dataset to advance video-based surveillance research for behavior recognition. It is from the well-known ITV's Emmerdale episodes involving dementia storyline.

1.1. Related work

Human action and behavior recognition has many potential applications including intelligent surveillance, assistive technologies, robotics and human-computer interaction. It is a fundamental and well-studied problem in computer vision with a long list of literature over the years [18, 6, 17]. Traditional approaches often depend on the hand-crafted feature extraction and representation (e.g. HOG [8] and SIFT [23]), hand-object interactions [37, 11], articulated pose [29, 24] and part-based/structured models [38, 9]. Many of these approaches explore the spatial configuration of body parts and hand-object interactions that often require body parts and/or object detector. Recently, major advances in CNN-based deep models [15, 25, 16] have challenged these approaches. These CNN models are trained and evaluated on very large and highly diverse datasets [29, 3] often consisting human-human, human-objects and human-animals interactions. In contrast, the targeted behavioral symptoms are often expressed via body language (e.g. gestures) and facial expression, and usually a hard problem for a machine to differentiate various symptoms shown by the same person. It is also known as *fine-grained* recognition.

Deep CNN models are comprised of multiple layers to learn representation of images/videos with multiple levels of abstractions through a hierarchical learning process [20]. Such models learn from very general (e.g. Gabor filters, edges, color blobs) to task-specific features as we move from first-layer to the last-layer [40]. Thus, these models are explored for TL in solving visual recognition tasks [12]. In TL, a *base* network is trained on a base dataset. Then, the learned features (e.g. weights) are adapted, or transferred to a second *target* network/model to be trained on a target dataset [40]. This would work if the learned features are task-independent, which means they are suitable for both base and target task. More recently, it has been shown that it is possible to obtain state-of-the-art results using TL [31, 12]. This suggests the layers of deep models do indeed learn features that are fairly general. In this paper, we explore strategies to strengthen this generalizability.

Automatic monitoring of the behavioral symptoms is often based on wearable sensors [5, 19, 7, 36]. In [7], Chikhaoui *et al.* have used Kinect and accelerometer to classify aggressive and agitated behavior using ensemble learning classifier. Whale *et al.* [36] used a smartphone app to collect context-sensitive information to monitor behavioral patterns that might be indicative of depressive symptoms. In [4], Chase *et al.* have used patients' movements to classify levels of motion that correlate with observed agitation. In [27], classification of aggressive actions (e.g. hitting, kicking, pushing and throwing) is carried out using skeleton joints obtained from a Kinect.

There has been very little progress in vision-based approach for unobtrusive monitoring of both aggressive and depressive behaviors in PwD. This could be due to the subjective behavior measurements, difficulty in measuring fine changes in appearance or even the lack of a public database. In this paper, we aim to address these issues by exploring the power of transfer learning (TL) through a novel simplistic approach and a new challenging dataset (Fig. 1) from ITV's special Emmerdale episodes involving dementia.

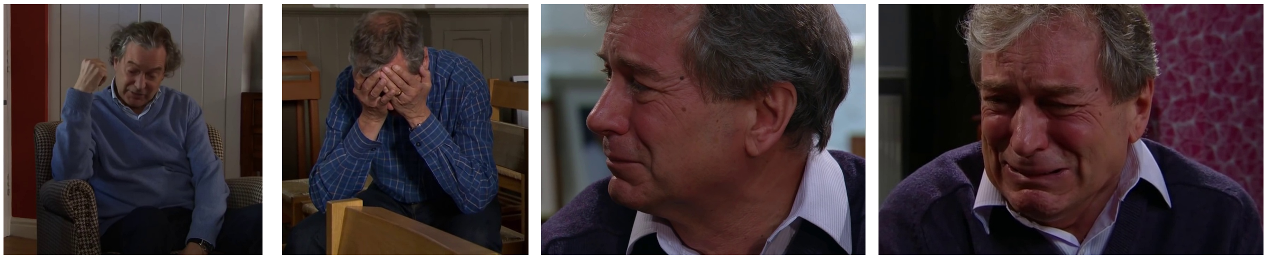
2. Proposed approach

We aim to recognize behavioral symptoms from still images by exploring the transferable CNN features. Recognizing behavior from still images is a challenging problem, is mainly due to the absence of temporal information. Furthermore, the objective is to maximize the use of TL to minimize resources (e.g. GPU) and computational time to train/validate the target model while still achieving competitive performance. The recognition based on still images would be computationally inexpensive. Therefore, it could be applied to real-world applications involving real-time monitoring for immediate intervention and/or support.

2.1. Convolutional Neural Network (CNN) features

State-of-the-art deep CNN models have achieved significant improvement in image recognition problems [35, 34, 32]. Given the complexity (number of parameters are in millions) of such models, it is necessary to train and evaluate them on large-scale image datasets [22, 30]. To train these models from scratch often requires multiple GPUs and can be a computationally expensive process. The training process could take days to weeks depending on the GPU memory/speed and size of the dataset. Thus, we benefited from the available pre-trained models on the ImageNet dataset [30] and explore the TL to solve the targeted fine-grained behavior recognition task. The ImageNet dataset consists of 1.2M natural images with 1K classes that includes people. Our target dataset (sec. 3.1) is smaller than the ImageNet and therefore more appropriate for TL.

Given the performance and relatively wider usages, we consider the VGG16 [32], Inception-V3 [35] and Inception ResNet-V2 [34] state-of-the-art deep models to extract CNN features. For a given input image, the forward pass extracts CNN features from the last layer before the softmax layer of the chosen pre-trained model. We use the default image size: 224×224 for VGG16 and 299×299 for ResNet-V2 and Inception-V3, resulting feature dimension of 4096 (VGG16), 1536 (V2) and 2048 (V3). In order to



(a) Upper body (b) Occluded face (c) Profile view (face) (d) Head and shoulder

Figure 1: Example images of depression behavioral symptom expressed in different ways

explore the appearance features representing different level of abstractions, as well as their suitability on the given task, we use three different extraction points (FC2, Block5 and Block4 pooling layers) in the VGG16 model. Thus, the CNN feature dimension is 4096, 25,088 and 25,088 for the FC2 (fully-connected), Block4 (B4) and Block5 (B5) pooling layers extraction points, respectively.

2.2. Hand-crafted feature (HOG)

In order to compare the performances of CNN features versus hand-crafted features, we use well-known Histogram of Oriented Gradient (HOG) feature [8]. The HOG parameters are set to default as in [8]. The resulting HOG feature dimension is 29,241 for an image size of 168 × 168.

2.3. SVM-based recognition of symptoms

We use the linear SVM to recognize behavioral symptoms from CNN features extracted using various deep models. To achieve this, we use LIBLINEAR [13] to solve:

$$\underset{\mathrm{w}}{\text{minimize}}\ \frac{1}{2}\|w^2\| + C\sum_i \max(1 - l_i w^T \mathbf{F_i}, 0)^2 \qquad (1)$$

Where $(\mathbf{F_i}, l_i)$ represents feature-label pair of i^{th} image. C is a penalty parameter and $\max(1 - l_i w^T \mathbf{F_i}, 0)^2$ is L_2-loss function. We use one-against-all strategy. The optimal value of C is very important for better performance and often decided through a grid search [13], which is computationally intensive. We overcome this by using the Bayesian optimization [33], in which the SVM learning algorithm in Eqn. 1 is modeled as a sample from a Gaussian process (GP). The GP attempts to find the optimal value of C in Eqn. 1 as few iterations as possible.

In order to evaluate our approach, we use the metrics (sec. 3.2) that require probabilistic outputs from a classifier. Linear SVM [13] provides outputs as class labels. Therefore, we use probability calibration method described in [26], which transforms linear SVM predictions to posterior probabilities by passing them through a sigmoid (also known as Platt calibration).

We evaluate the proposed method by combining various CNN features. We use feature-level fusion by concatenating various features into a single feature vector, which is then used by the SVM for classification.

2.4. Fine-tuning deep models on the target dataset

We retrain the state-of-the-art VGG16 [32], Inception-V3 [35] and Inception ResNet-V2 [34] models on the target dementia dataset by applying TL. This is inspired by the findings in [40], which suggests that transferred weights perform better in comparison to random weights for both frozen (i.e. do not change during training) and fine-tuning on a new dataset. During fine-tuning, all layers are initialized with pre-trained weights except the softmax layer (random initialization) due to the different number of classes. We use the default settings (e.g. pre-processing step, image size, data augmentation) except the batch size, which is set to 32 to fit into 8GB GPU (NVIDIA Quadro M4000). We use stochastic gradient descent (SGD) optimizer to minimize the categorical cross entropy $e_i = -\sum_c l_{i,c} log(p_{i,c})$, where p are the predictions, l are the true labels, i denote the training images and c represent the behavior categories. In our experiments, we set *leaning rate* to 10^{-4}, *momentum* to 0.95 and *decay* to 0.005.

3. Experiments, evaluations and discussion

This section describes the new dataset, experimentation details and reports the performances.

3.1. Dataset

The new dementia dataset is a collection of still frames from the ITV soap opera Emmerdale episodes focusing on the dementia storyline. These episodes are approved by the Alzheimer's Society to raise awareness of the issues surrounding dementia. The dataset is created using 57 short YouTube clips and 13 episodes (each of duration ∼35 minutes) featuring the dementia character (*Ashley Thomas*). The images are categorized into 4 labels; *aggressive*, *depressive*, *happy* and *neutral*. The dataset consists of 65,082 images (aggressive: 13,535; depressive: 21,075; happy: 13,841 and neutral: 16,631). Most of the images are cropped to focus on the main character *Ashley Thomas*. The data is split into ∼20% test and ∼80% train data. The split is done in such a way to ensure that very similar frames are kept in either the training or testing set. The dataset information is available at: https://computing.edgehill.ac.uk/~abehera/.

3.2. Evaluation metrics

We use the standard metric of accuracy (ACC) and average precision (AP). ACC gives the percentage of correct predictions and assigns equal cost to false positives and false negatives. Whereas, AP summarizes precision-recall curve. We also compute log loss $logLoss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{4} y_{i,c}log(p_{i,c})$, where $p_{i,c}$ is the prediction probability of class c given i^{th} test image and $y_{i,c}$ is actual probability (0 or 1). N is the total number of test images. The log loss quantifies the accuracy of a classifier by penalizing the confident false classifications. An ideal classifier would have zero log loss. We also consider the top-2 ACC.

3.3. Discussion

The performance of individual CNN features using linear SVM is shown in Table 1. It is evident that the CNN features perform better than the HOG [8] hand-crafted feature. CNN feature extracted from the B5 gives better performance than the FC2 and B4 of VGG16 [32]. This shows more than one extraction point should be considered while using CNN features extracted via TL. The top-2 (ACC-2) accuracy using Inception-V3 is better than the rest. By looking at the log loss (smaller is better), there is not much difference in classifier's confidence in making decision. The classifiers using VGG16 B5, Inception-V3 and Inception ResNet-V2 features have better confidence in making decisions than the rest (Table 1).

Table 2 shows performance by combining (feature-level fusion) two or three different types of CNN features. Given the individual performance (Table 1) and sizes, B4 and HOG features are not considered for fusion. A general observation is that the overall performance improves with various combinations. However, none of these performance is better than the VGG16 B5 (Table 1) alone. For example, when B5 is combined with other features, the accuracy is better than the individual accuracy except B5 (50.8%). A similar trend is also observed in combining three features. The performance of classifier-level fusion involving three different CNN models is presented in Table 3. The overall performance is lower than the feature-level fusion (Table 2).

Features	ACC	ACC-2	AP	Log loss
Inception-V3	41.9	**73.8**	42.8	1.23
Inception ResNet-V2	49.7	67.6	46.5	1.23
VGG16 B4	41.9	69.3	43.9	1.26
VGG16 B5	**50.8**	71.1	**51.6**	**1.23**
VGG16 FC2	45.0	63.7	42.9	1.26
HOG	31.5	61.2	29.6	1.35

Table 1: Performance of various CNN and hand-crafted HOG features in percentage (except log loss) using SVM.

Features	ACC	ACC-2	AP	Log loss
V3 + V2	49.9	69.0	46.7	**1.24**
V3 + B5	50.4	71.6	49.0	1.25
V3 + FC2	44.9	69.7	48.8	1.25
V2 + B5	**50.7**	71.8	**49.9**	1.25
V2 + FC2	46.2	70.0	48.0	1.26
B5 + FC2	50.3	**72.1**	49.3	1.25
V3 + V2 + B5	50.2	71.9	48.9	1.26
V3 + FC2 + B5	**50.5**	**72.2**	48.9	1.25
V2 + V3 + FC2	46.9	70.3	**51.0**	**1.24**
V2 + B5 + FC2	50.4	71.9	48.9	1.26

Table 2: Performance of feature-level fusion (concatenation) in percentage using SVM with various combination. $B5$ → VGG16 [32] Block5 pooling layer, $B4$ → VGG16 [32] Block4 pooling layer, $FC2$ → VGG16 [32] FC2 layer, V_2 → Inception ResNet-V2 [34] and V_3 → Inception-V3 [35].

The performance of the state-of-the-art models fine-tuned on the target dataset is presented in Table 4. The models are trained for 50 epochs using the TL approach as described in section 2.4. It is clear that the performance of Inception-V3 (ACC: 49.6%) model is better than the rest. An interesting observation is that CNN feature from VGG16 B5 outperformed these deep models. This shows CNN features with appropriate level of abstraction, when combined with basic classifier such as linear SVM could give better performance than the re-training/fine-tuning these models on a target dataset. Re-training/fine-tuning often requires GPUs and optimization of various hyper-parameters. Thus, it can be time consuming and a computationally expensive process.

We have also experimented with CNN features extracted from the above-mentioned fine-tuned models. In this sce-

SVM Kernel	ACC	ACC-2	AP	Log loss
Linear	43.0	63.9	**46.4**	**3.55**
RBF	**45.9**	**70.0**	45.1	9.24

Table 3: Performance of classifier-level fusion in which the outputs from individual SVM trained on VGG16, ResNet-V2 and Inception-V3 features are fed into another SVM with linear and RBF kernel.

Features	ACC	ACC-2	AP	Log loss
Inception-V3	**49.6**	**73.1**	**54.6**	**1.22**
Inception ResNet-V2	45.1	67.8	45.0	1.24
VGG16	45.6	69.3	50.3	4.79

Table 4: Performance of CNN models in percentage (except log loss) fine-tuned (transfer learning) on the target dataset.

Features	ACC	ACC-2	AP	Log loss
Inception-V3	50.4	**74.9**	**54.4**	1.39
Inception ResNet-V2	37.4	69.2	43.4	1.43
VGG16 B4	45.6	70.4	46.6	1.23
VGG16 B5	**52.4**	73.0	52.4	**1.16**
VGG16 FC2	48.2	69.4	50.2	2.01

Table 5: SVM performance in percentage (except log loss) using various CNN features extracted from the respective models, which are fine-tuned on the target dataset.

nario, instead of pre-trained models, we use the respective fine-tuned models to extract CNN features. The performance is presented in Table 5. The performance is better than the fine-tuned models in Table 4 and CNN features from pre-trained models in Table 1 with the exception of Inception ResNet-V2. The VGG16 B5 performed best in comparison to the rest. In comparison to the performance in Table 1, the most improved (ACC: 8.5%) model is the Inception-V3. The other notable observation is that using feature-level fusion the performance is improved while using the fine-tuned models (Table 6 vs Table 5) in comparison to the pre-trained models (Table 2 vs Table 1).

The confusion matrices for various experiments are shown in Fig. 2. The performance is very good given the challenging dataset, which consists of fine changes in appearance representing different behavioral symptoms. For example, the symptom *happy* is often noticed via facial expression. It is difficult for machine to recognize since the visible area representing the face is often small and heavily depends on the head orientation (*i.e*. frontal vs profile view). This could be linked to the low performance of *happy* and *neutral* symptoms (Fig. 2). The high accuracy of *aggressive* and *depressive* behaviors could be linked to the visible scale range *i.e*. when the character is expressing such symptoms, the camera is often zoomed to the upper-body. Moreover, these expressions are often accompanied with other body-languages like hand gestures and hand-over-face (Fig. 1).

4. Conclusions

We have presented the challenges in video-based unobtrusive monitoring of the behavioral symptoms in PwD. In order to recognize such symptoms in video frames, we proposed a novel method that explored the power of transfer learning. The proposed approach is simple and used a data-driven approach that applied different level of abstractions using the state-of-the-art deep CNN models. We have introduced a new dataset from ITV's Emmerdale episodes involving the dementia storyline, which is described as a "realistic portrayal" of the condition by the Alzheimer's Society[1]. We believe this will help advance the field of video surveillance focusing on behavioral symptoms in PwD.

Features	ACC	ACC-2	AP	Log loss
V3 + V2	51.0	**75.1**	**55.7**	1.37
V3 + B5	52.9	72.9	52.9	1.16
V3 + FC2	48.4	69.4	50.8	2.01
V2 + B5	**53.0**	73.1	52.6	**1.16**
V2 + FC2	48.0	69.3	50.4	2.01
B5 + FC2	52.9	72.9	52.7	1.23
V3 + V2 + B5	**54.7**	73.1	**52.6**	**1.16**
V3 + FC2 + B5	53.8	72.8	52.5	1.17
V2 + V3 + FC2	47.9	69.3	50.1	2.00
V2 + B5 + FC2	54.2	**73.3**	52.6	1.16

Table 6: SVM performance in percentage (except log loss) by fusing various CNN features extracted from the respective models, which are fine-tuned on the target dataset.

Acknowledgments: This research was supported by the Edge Hill University-funded Student Opportunity Fund (SOF) and Research Investment Fund (RIF).

References

[1] Alzheimer's Society. About depression and alzheimers disease, 2010. Factsheet 444.

[2] Alzheimer's Society. Aggressive behaviour, August 2017. Factsheet 509LP.

[3] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images. In *IEEE ICCV*, pages 1017–1025, 2015.

[4] J. G. Chase et al. Quantifying agitation in sedated icu patients using digital imaging. *Computer methods and programs in biomedicine*, 76(2):131–141, 2004.

[5] M. Chen, S. Gonzalez, A. Vasilakos, H. Cao, and V. C. Leung. Body area networks: A survey. *Mobile networks and applications*, 16(2):171–193, 2011.

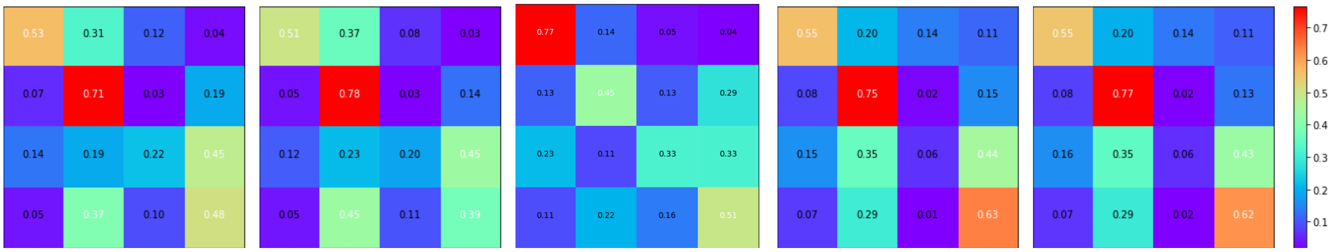
[6] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles. Advances in human action recognition: A survey. *arXiv preprint arXiv:1501.05964*, 2015.

[7] B. Chikhaoui, B. Ye, and A. Mihailidis. Ensemble learning-based algorithms for aggressive and agitated behavior recognition. In *Ubiquitous Comp. and Ambient Intelligence*, 2016.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR*, pages 886–893, 2005.

[9] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010.

[10] A. K. Desai and G. T. Grossberg. Recognition and management of behavioral disturbances in dementia. *Primary care companion to the jrnl of clini. psych.*, 3(3):93–109, 2001.



(a) B5 feature and SVM (b) B5 + V2 feature and SVM (c) Fine-tuned Inception-V3 (d) Fine-tuned B5 and SVM (e) Fine-tuned B5 + V2 and SVM

Figure 2: Confusion matrices (y-axis, top-to-bottom: actual and x-axis, left-to-right: predicted) for recognizing behavioral symptoms (1-*aggressive*, 2-*depressive*, 3-*happy* and 4-*neutral*). The performance of *depressive* is the best (except in c) and *happy* symptom is the worst.

[11] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. In *IEEE CVPR workshops*, pages 9–16, 2010.

[12] J. Donahue et al. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, volume 32, pages 647–655, 2014.

[13] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

[14] Z. Fei et al. A survey of the state-of-the-art techniques for cognitive impairment detection in the elderly. In *Adv. Comput. Methods in Life Sys. Modeling and Simulation*, pages 143–161, 2017.

[15] R. Girdhar and D. Ramanan. Attentional pooling for action recognition. In *Advances in NIPS*, pages 33–44, 2017.

[16] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r*cnn. In *ICCV*, pages 1080–1088, 2015.

[17] G. Guo and A. Lai. A survey on still image based human action recognition. *Patt. Recog.*, 47(10):3343 – 3361, 2014.

[18] S. Herath, M. Harandi, and F. Porikli. Going deeper into action recognition: A survey. *IVC*, 60:4–21, 2017.

[19] S. S. Khan, B. Ye, B. Taati, and A. Mihailidis. Detecting agitation and aggression in people with dementia using sensorsa systematic review. *Alzheimer's & Dementia*, 2018.

[20] Y. LeCun, Y. Bengio, and G. E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[21] M. Leo, G. Medioni, M. Trivedi, T. Kanade, and G. Farinella. Computer vision for assistive technologies. *Computer Vision and Image Understanding*, 154:1 – 15, 2017.

[22] T.-Y. Lin et al. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, Nov. 2004.

[24] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *IEEE CVPR*, pages 3177–3184, 2011.

[25] A. Mallya and S. Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *ECCV*, pages 414–428, 2016.

[26] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proc. of the 22nd Int'l Conf. on Machine Learning (ICML)*, pages 625–632, 2005.

[27] S. Nirjon et al. Kintense: A robust, accurate, real-time and evolving system for detecting aggressive actions from streaming 3d skeleton data. In *IEEE Pervasive Computing and Communications*, pages 2–10, 2014.

[28] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers. A review of wearable sensors and systems with application in rehabilitation. *Journl of neuroeng. and rehab.*, 9(1):21, 2012.

[29] L. Pishchulin, M. Andriluka, and B. Schiele. Fine-grained activity recognition with holistic and pose based features. In *German Conf. on Pattern Recog.*, pages 678–689, 2014.

[30] O. Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015.

[31] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *ICLR*, 2014.

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[33] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in NIPS*, pages 2951–2959, 2012.

[34] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284, 2017.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE CVPR*, pages 2818–2826, 2016.

[36] F. Wahle, T. Kowatsch, E. Fleisch, M. Rufer, and S. Weidt. Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR mHealth and uHealth*, 4(3), 2016.

[37] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *IEEE CVPR*, pages 17–24, 2010.

[38] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *IEEE ICCV*, pages 1331–1338, 2011.

[39] M. Yefimova and D. L. Woods. Using sensor technology to monitor disruptive behavior of persons with dementia. In *AAAI Fall Symposium: AI for Gerontechnology*, 2012.

[40] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in NIPS*, pages 3320–3328, 2014.