PAPER • OPEN ACCESS

Deep learning-based artificial vision for grasp classification in myoelectric hands

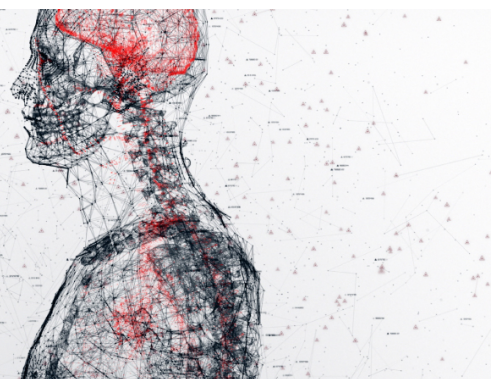View the article online for updates and enhancements.

physicsworld

AI in medical physics week
20–24 June 2022

Join live presentations from leading experts in the field of AI in medical physics.

physicsworld.com/medical-physics

Deep learning-based artificial vision for grasp classification in myoelectric hands

Ghazal Ghazaei[1], Ali Alameer[1], Patrick Degenaar[1,2], Graham Morgan[3] and Kianoush Nazarpour[1,2]

[1] School of Electrical and Electronic Engineering, Newcastle University, Newcastle-upon-Tyne NE1 7RU, United Kingdom
[2] Institute of Neuroscience, Newcastle University, Newcastle-upon-Tyne NE2 4HH, United Kingdom
[3] School of Computing Science, Newcastle University, Newcastle-upon-Tyne NE1 7RU, United Kingdom

E-mail: G.Ghazaei1@newcastle.ac.uk and Kianoush.Nazarpour@newcastle.ac.uk

Abstract

Objective. Computer vision-based assistive technology solutions can revolutionise the quality of care for people with sensorimotor disorders. The goal of this work was to enable trans-radial amputees to use a simple, yet efficient, computer vision system to grasp and move common household objects with a two-channel myoelectric prosthetic hand. *Approach*. We developed a deep learning-based artificial vision system to augment the grasp functionality of a commercial prosthesis. Our main conceptual novelty is that we classify objects with regards to the grasp pattern without explicitly identifying them or measuring their dimensions. A convolutional neural network (CNN) structure was trained with images of over 500 graspable objects. For each object, 72 images, at 5° intervals, were available. Objects were categorised into four grasp classes, namely: pinch, tripod, palmar wrist neutral and palmar wrist pronated. The CNN setting was first tuned and tested offline and then in realtime with objects or object views that were not included in the training set. *Main results*. The classification accuracy in the offline tests reached 85% for the seen and 75% for the novel objects; reflecting the generalisability of grasp classification. We then implemented the proposed framework in realtime on a standard laptop computer and achieved an overall score of 84% in classifying a set of novel as well as seen but randomly-rotated objects. Finally, the system was tested with two trans-radial amputee volunteers controlling an i-limb Ultra™ prosthetic hand and a motion control™ prosthetic wrist; augmented with a webcam. After training, subjects successfully picked up and moved the target objects with an overall success of up to 88%. In addition, we show that with training, subjects' performance improved in terms of time required to accomplish a block of 24 trials despite a decreasing level of visual feedback. *Significance*. The proposed design constitutes a substantial conceptual improvement for the control of multi-functional prosthetic hands. We show for the first time that deep-learning based computer vision systems can enhance the grip functionality of myoelectric hands considerably.

Keywords: myoelectric hand prosthesis, convolutional neural network, grasp classification

S Supplementary material for this article is available online

(Some figures may appear in colour only in the online journal)

1. Introduction

Prosthetic hands can provide a route to functional rehabilitation of upper-limb amputees and people with congenital motor deficit. According to recent statistics, in UK alone, there are 473 new upper-limb (133 trans-radial) referral every year; of which, 245 are in the age range of 15 and 54 years old [1]. Lifetime care for this group can be remarkably expensive. Trauma is the most prevalent cause of limb loss at ∼30% [1]. In the US, there are around 500 k upper-limb amputees [2]. Advanced prosthetic hands can dramatically improve users' quality of life by enabling them to carry out daily living activities.

Current commercial prosthetic hands are typically controlled via the myoelectric signals, that is the electrical activity of muscles recorded from the skin surface of the stump [3, 4]. Despite considerable technical advances and improvements in the mechanical features, e.g. size and weight, of the prosthetic hands, the control of these systems is still limited to one or two degrees of freedom [4, 5]. In addition, the process of switching a prosthetic hand into an appropriate grip mode, e.g. pinch, can be cumbersome or would require an ad-hoc solution, such as using a mobile application[4] or via an Electrocutaneous menu [6].

For several decades, research on prosthetic control has focused on myoelectric pattern recognition [3, 4]. Classification and proportional control of myoelectric signals has been extensively studied for discrete decoding of wrist and elbow movements [7–14], grasp type [15–17], as well as individuated finger movements [18], with accuracies as high as 90% [19] in amputee subjects. Although reasonable classification accuracies are gained, there is still a considerable gap between the laboratory-based research and the widespread clinical use of pattern recognition-based systems. Lack of robustness, number and movement of the electrodes as well as modulation of the electromyogram (EMG) signal activation patterns with varying force and orientation of the arm may be the main reasons [4, 14, 20]. To become fully integrated into an amputee's sensorimotor repertoire, the performance of hand prostheses must still improve greatly [21–24]. The COAPT system is the first commercial myoelectric controller unit to employ pattern recognition[5].

As intermediate solutions, alternative modalities have been adopted to replace or augment the EMG signals. Skin movement analysis via accelerometry signals [25, 26], force myography [27], use of radio-frequency identification (RFID) tags [28], arm movement trajectory and inertial measurement (e.g. i-mo™) and computer vision [29–34] are some examples.

Specifically, in the case of using computer vision, it was shown that object shapes can be quantised such that appropriate grasp types and sizes can be determined. Došen *et al* [29, 30] demonstrated a dexterous hand with an integrated vision based control system. The user controlled the prosthesis hand and the activation of the camera with myoelectric signals. A simple object detection method was used, in conjunction with distance information, estimated via ultrasound. This structure allowed them to approximate the size of the object of interest. The calculated size was then introduced to a rule-based reasoning algorithm to select the appropriate grasp accordingly.

They achieved 84% accuracy in estimating the grasp type and size for a limited set of 13 objects (93%, grasp only). Acquiring such level of accuracy, each trial took on average ∼4 s, on a dual-core 2 GHz PC, since classification of 10 consecutive snapshots was required for each decision. Marković *et al* [31] demonstrated a semi-autonomous control mechanism in which stereo-vision provided depth information. In addition, their solution offered artificial proprioceptive feedback, via visual feedback to the user, about the grip aperture size by using augmented reality (AR). They incorporated sophisticated algorithms for image segmentation, 3-dimensional point cloud generation and geometrical model fitting. These algorithms however used a similar rule-based model that was proposed earlier by Došen *et al* [29, 30]. With such improvements, the process of identifying the object size and the appropriate grasp became significantly faster, about 1 s, on an Intel i5 core (2.73 GHz) laptop with 8 GB of RAM. They achieved an overall accuracy of 81% for the successful accomplishment of the task (∼94% in grasp identification). However, without the AR feedback, this accuracy dropped to 73%. In [29–31], authors included four grasp types, namely, palmar, lateral, tri-digit (here: tripod) and bi-digit (here: pinch).

Marković *et al* [33] further exploited a data fusion technique to control a prosthetic hand. A plethora of modalities, namely, myoelectric recording, computer vision, inertial measurements and embedded prosthesis sensors (position and force) were utilised to provide realtime simultaneous, proportional and semi-autonomous control. The shape, the size and the orientation of objects were estimated with RGB-D imaging and integrated with prosthesis orientation and user behaviour via inertial sensing. Such a sophisticated architecture led to less than 1% cumulative trial failure rate. This setting was integrated into a prosthetic wrist, but only palmar and lateral grasps were considered.

Computer vision has been widely used in *robotic* grasp and object manipulation [35–38]. Saxena *et al* [35] pioneered the field by providing the capability of grasping novel (unseen) objects for robotic hands by utilising a stereo camera. Without building a 3-dimensional model, they estimated the 3-dimensional location of the best grasp by triangulation. The grasp location estimator algorithm was trained on synthetic images in a supervised learning regime. Kootstra *et al* [36] developed an *early* cognitive vision architecture for grasping unknown objects. Without any segmentation or preprocessing, they were able to generate two- and three-finger grasps based on contours and surface structure provided by stereo cameras. With the advancement of the deep learning structures [39], robotic grasp research has been radically upgraded. For instance, Lenz *et al* [37] introduced RGB-D images to a two-step cascade deep learning system. Given the image of an object to grasp, firstly a small deep network determined the suitable grasping points for the object; based on its position, size and orientation. Then, a second network was trained to pick the best candidate among the grasping spots that were identified by the first network. Group regularisation was utilised to balance learning with respect to information extracted from different modalities, such as the colour of the object, depth and surface normals. Similarly, Kopicki *et al* [38] provided a one-shot learning

---

Table 1. A list of current prosthetic and robotic hands that use vision. The letters 'P' and 'R' in the Field column denote prosthetics and robotics, respectively. In the top three rows, the shown success rates reflect the identification of the correct grasp types only.

Related work	Field	Success (%)	Time(s)	Hand
[29, 30]	P	93	~4	CyberHand
[31]	P	94	~1	SmartHand
[33]	P	~99	0.75	Michelangelo hand
[35]	R	87	1.2	2-finger gripper
[36]	R	20–60	N/A	2/3-finger gripper
[37]	R	93.7	13.5	2-finger gripper
[38]	R	77.8	13–24	Boris hand

mechanism for recognising the most appropriate grasp for novel objects. They generated thousands of grasp candidates for images taken by a depth camera and optimised the combination of two learned model types: a contact model and a hand-configuration model. Table 1 shows a summary of structures that utilised vision in prosthetic and robotic applications.

We set out to translate the advances in deep learning in the robotics and computer vision research for control of hand prostheses. Benefiting from the flexibility that a deep learning structure offers, we developed an inexpensive vision-based system suitable for use in artificial hands. This solution can identify the appropriate *grasp type* for objects according to a learned abstract representation of the object rather than the explicitly-measured object dimensions. This key concept is illustrated in figure 1. In this way, objects are not classified based on the object category or identity, but based on the suitable grasp pattern. A key question would therefore be whether this deep learning-based approach generalises to unseen objects. We predict that a deep network trained for grasp recognition can extract high-level and grasp-related features from objects and discard other unnecessary details. These features could include object size and orientation. This approach is therefore conceptually different from object recognition in which object details matter.

To learn this abstract representation, we use a convolutional neural network (CNN) architecture [39]. There is mounting evidence that CNN-based structures can learn and classify visual patterns efficiently if provided with a large amount of training (labelled) samples [40–45]. The components of the CNN structure, namely, local connectivity, parameter sharing and pooling, make it reasonably invariant against object shift, scale and distortion. These features make the CNN structure a suitable candidate for upper-limb prosthetics applications. We therefore trained a CNN structure to identify the appropriate grasp for a database of household objects. The CNN structure, or in fact any other supervised learning architecture in which there exists a set of predefined labels, lack the ability of generalisation to novel objects that do not belong to any defined output object categories. Therefore during testing, unseen objects will be misclassified to one of the existing classes. However, identification of novel objects is crucial in prosthetic applications; since in everyday life people effortlessly pick up a variety of objects that they have never seen before. Moreover, the number of the categories of household

objects can be excessively large, making object identification for grasp selection impractical.

2. Methods

In this section, we give a detailed description of the equipment and methods that we used offline and in the realtime experiments, both in computer-based tests and when amputee users controlled the prosthesis. To enhance clarity and in the interest of brevity, we merge the description of the methods that were common in all experiments.

2.1. Image databases

To train the CNN structure, we used the Amsterdam library of object images (ALOI) [46]. The ALOI database offers a rich set of the images of household objects. To enable real-time testing, we augmented the ALOI dataset by our dataset which we call Newcastle Grasp Library (made freely available online, see Acknowledgements). In the following, we describe both image libraries.

2.1.1. Amsterdam library of object images (ALOI). The ALOI database [46] includes the images of 1000 common objects. Within this library, 250 objects have been photographed at a second zoom rate. We discarded these 250 objects. For each of the remaining 750 objects, the database includes 72 pictures, taken at 5° intervals against a black background. The camera was at 124.5 cm distance and 30 cm altitude from the objects. The camera resolution was 768×576 pixels. We *subjectively* selected 473 of the objects in four different classes of pinch, tripod, palmar wrist neutral and palmar wrist pronated. Other objects were either not graspable or could be picked with more than one grasp type. All images were first converted to grey-scale. They were then downsampled to a resolution of 48×36 pixels; using the *imresize* function in MATLAB®.

2.1.2. Newcastle grasp library. Access to the same objects that were used to create the ALOI database was not possible. Therefore, to enable realtime analysis, 71 objects in four grasp classes were selected for photography. We synchronised a Crayfish 55 turntable (Seabass, UK) with a Canon Kiss X4 DSLR camera (resolution 18 Megapixel, 5184×3456 pixels) to take 72 pictures from each object (at 5° intervals) against a black background. Table 2 indicates the number of objects in each grasp group that we used for further analysis.

To ensure object size is taken into account we positioned the camera at a fixed distance from objects when collecting the images. The distance between the camera and the object was 60 cm and the webcam was 15 cm higher than object. With this setting we could achieve images of objects that were comparable in size with those available in the ALOI database. All images were converted to grey-scale and downsampled to a resolution of 48×36 pixels to train the CNN setting.

Figure 2(A) represents some of the objects we selected from the ALOI database. Figure 2(B) shows all the additional objects included in the Newcastle Grasp Library. A list of all the objects that we chose and the corresponding grip

A)

B)

Figure 1. Object versus grasp recognition. (A) Object recognition. (B) Grasp recognition.

types are reported as supplementary material (stacks.iop.org/JNE/14/036025/mmedia).

2.2. Feature extraction—convolutional neural network (CNN)

As mentioned earlier, each image I was first converted to grey-scale and was downsampled to an $N = 36$ by $M = 48$ image. It was then passed through Gaussian and median filtering for noise removal and smoothing. Empirically, we found that image normalisation, prior to the CNN setting, improved the final accuracy. Therefore, each image was normalised according to

$$I_{\text{normalised}} = \frac{(I - \mu_I)}{\sigma_I} \qquad (1)$$

where

$$\mu_I = \frac{1}{N + M} \sum_{n=1}^{N} \sum_{m=1}^{M} I_{n,m} \qquad (2)$$

and

Table 2. The number of objects included in the ALOI and Newcastle databases in each grasp group.

Grasp type \ Database	ALOI	Newcastle
Pinch	90	19
Tripod	163	11
Palmar wrist neutral	83	30
Palmar wrist pronated	137	11
Overall	473	71

$$\sigma_I = \sqrt{\frac{1}{N + M} \sum_{n=1}^{N} \sum_{m=1}^{M} (I_{n,m} - \mu_I)^2}. \qquad (3)$$

In the above equations, $I_{n,m}$ denotes the intensity of pixel (n, m).

For classification of images into grasp groups, we examined two CNN architectures: a one-layer and a two-layer, and explored the trade-off between accuracy, generalisability and computational complexity. We first explain briefly the setting of the developed CNN structure. In the following, all equations are presented in the vectorised format.

A)

Pinch | Tripod | Palmar Wrist Neutral | Palmar Wrist Pronated

B)

Pinch | Tripod | Palmar Wrist Neutral | Palmar Wrist Pronated
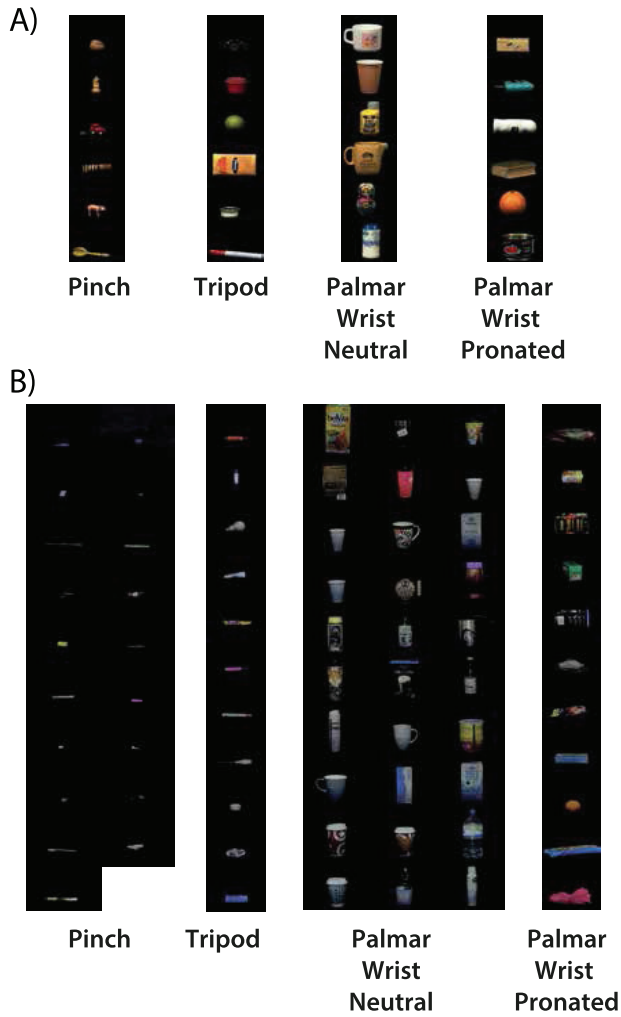
Figure 2. The databases of objects used in this paper and their corresponding grasp type. (A) A small subset of objects in the ALOI database; (B) All objects in the Newcastle Grasp Library. All images were converted to grey-scale and downsampled before further analysis.

Assume that there are m^l input maps of size $R^l \times U^l$ in each of the CNN layers l, with m^0 denoting the number of images in the 0-th layer. k^l features are extracted at each layer by convolution with $C^l \times D^l$ ($C^l < R^l, D^l < U^l$) kernels according to

$$\mathbf{Z}_{ij}^l = (\mathbf{X}_{ij}^{l-1} * \mathbf{W}_j^l + \mathbf{b}_j^l) \tag{4}$$

$$\mathbf{X}_{ij}^l = a(\mathbf{Z}_{ij}^l) \tag{5}$$

where $\mathbf{Z}_{ij}^l$ is a $(R^l - C^l + 1) \times (U^l - D^l + 1)$ matrix resulted from convolving the i-th input map from the $(l-1)$-th layer ($\mathbf{X}_{ij}^{l-1}$) and the j-th kernel in the l-th layer ($\mathbf{W}_j^l$) and adding the bias $\mathbf{b}_j^l$. The output of layer l is then calculated by element-wise application of the activation function $a(\cdot)$. In the above equation, the asterisk sign $*$ refers to a *valid* convolution, that is, a convolution performed inside the image borders. Finally, $i = 1, 2, \ldots, m^l$, $j = 1, 2, \ldots, k^l$ and $l = 0, 1, 2, \ldots, L$.

We tested a range of activation functions, namely, the logistic, hyperbolic tangent and rectified linear unit (ReLU)

functions. We empirically found that the ReLU function results in the highest performance and hence we used it in this study. The ReLU activation function $a(\cdot)$ can be written as

$$a(z_{r,u}) = \max(0, z_{r,u}) \tag{6}$$

where $z_{r,u}$ denotes an element of $\mathbf{Z}$ [47].

Our one-layer CNN comprised one convolution (C_1) and one sub-sampling (S_1) sub-layers. In the two-layer CNN architecture, however, we had two convolution C_1 and C_2 and one sub-sampling S_2 stages, of which the latter two were in the second layer.

In both CNN settings, we used five kernels ($\mathbf{W}_j^l$, $j = 1, \cdots, 5$) of size 5×5 and the resultant feature maps were sub-sampled by max-pooling [48] by a factor of two. We applied the max-pooling operation to ensure salient elements in each feature map are retained. With the max-pooling operation, each sub-region is replaced with the maximum value of that sub-region. Figure 3 illustrates the two-layer CNN setting with all details in terms of kernels and dimensions that we used in this study. This setting was adopted after a large number of empirical testing with different number of layers and filters, filter and pooling sizes and activation functions. Between all, we selected the setting in figure 3 that maximised the overall classification performance, specially in identifying the appropriate grasp for novel objects.

2.3. Classifier—softmax regression

Following the proposed CNN-based feature extraction, for classification, we used Softmax (or multi-nomial logistic) regression [49, 50].

Having m examples $\mathbf{x}^{(i)}$ and their corresponding class labels $y^{(i)}$ in a training set as $\{(\mathbf{x}^{(1)}, y^{(1)}), \cdots, (\mathbf{x}^{(m)}, y^{(m)})\}$, we estimate the probability $P(y = g|\mathbf{X})$ for $g = \{1, \cdots, G\}$ and $G > 2$. The matrix $\mathbf{X}$ has sample $\mathbf{x}^{(i)}$ in its i-th column. The matrix of model parameters $\boldsymbol{\theta}$ can be estimated by optimising the following cost function where $1\{\cdot\}$ is the 'indicator function', that is, $1\{\text{a true statement}\} = 1$ and $1\{\text{a true statement}\} = 0$ [51].

$$J(\boldsymbol{\Theta}) = \frac{-1}{m} \left[\sum_{i=1}^{m} \sum_{g=1}^{G} 1\{y^{(i)} = g\} \log \frac{e^{\boldsymbol{\theta}^{(g)T}\mathbf{x}^{(i)}}}{\sum_{j=1}^{G} e^{\boldsymbol{\theta}^{(j)T}\mathbf{x}^{(i)}}} \right] \tag{7}$$

where $(\cdot)^T$ denotes the vector transpose operation.

Training was carried out through back propagation using the mini-batch momentum gradient descent algorithm [52] for optimising the learned filters within each iteration. We avoided over-fitting by using Tikhonov regularisation in the final cost function during training the CNN structure where the matrix $\mathbf{W}_j^l$ in the last layer is optimised.

2.4. Cross-validation

To verify the generalisability and robustness of grasp classification, we examined two forms of cross-validation: within- and between-object cross-validations. In the following we introduce and provide the rationale for using them. Both of the CNN settings (one- or two-layers) were tested in both of the below cross-validations schemes.

The implemented two-layer CNN
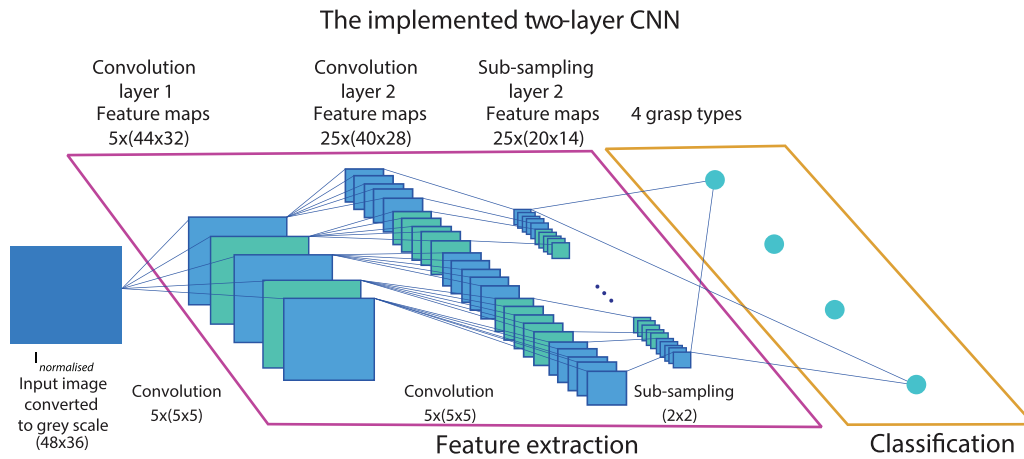
Figure 3. The implemented two-layer CNN architecture.

2.4.1. Within-object cross-validation (WOC). Firstly, we evaluated the ability of the proposed structure in classifying previously *seen* objects. The training set included 90% (65 of 72) of the views for each object in each grasp class. The remaining 10% of the views for each object were allocated to the testing set. We randomly selected 10 different training and testing sets to quantify the sensitivity of the classifier to the choice of views. Figure 4 illustrates an example of splitting images of one object into the training and testing sets.

2.4.2. Between-object cross-validation (BOC). To be able to identify the appropriate grasps for *unseen* objects, we carried out the BOC test. In the BOC scheme, an object and its views were either wholly *seen* or *unseen*.

For the ALOI database, the training set included ~90% of all the object categories in all grasp groups with all of their different poses; for instance all 124 objects of the 'palmar wrist pronated' class with all their 72 poses were selected for training. The remaining ~10% of the object categories were allocated to the testing set, that is, 13 objects in this class.

An example for random selection of 4 objects from the Newcastle Grasp Library in the 'palmar wrist pronated' class for the test set is illustrated in figure 5. The above procedure was repeated 10 times independently. Table 3 reports the exact number of objects selected for training and testing from each database in the BOC test.

2.5. Statistical analysis

A two-way repeated-measures ANOVA was conducted that examined the main effects of cross-validation type (BOC versus WOC) and number of layers (1 versus 2) in the CNN structure on the offline classification results. In this analysis, each of the 10 folds of cross-validation was treated as an independent sample. In the realtime experiments with amputee subjects, we compared the average block accomplishment times in blocks 1 and 6 with a paired t-test, for each participant independently. All tests were performed in SPSS® 22.
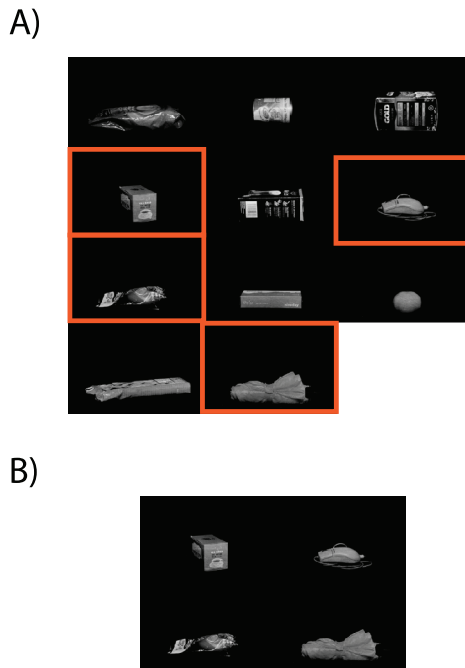
Within-Object Cross-Validation
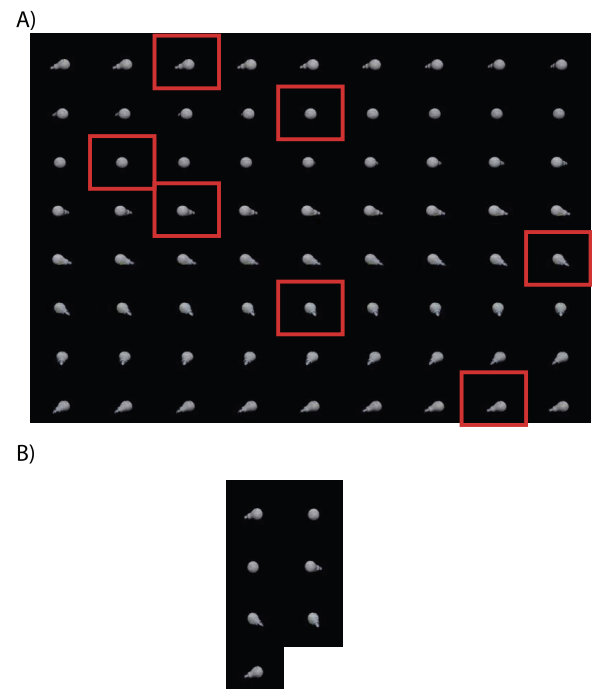
Figure 4. Within-object cross-validation. The object is a plastic light bulb and belongs to the Newcastle Grasp Library. The testing set (B) is a randomly selected subset (10%) of all views available in (A) (shown in red boxes). This figure shows one example of 10 cross-validation folds. All original images were downsampled before further analysis. (A) Training set. (B) Testing set.

2.6. Computer-based realtime performance analysis

We implemented the introduced deep-learning based system in realtime. We carried out the realtime experiments with the learned CNN parameters of the BOC setting. This was because in the real-life cases, it is likely that novel objects are encountered.

We deliberately included this stage before real-time experiment with amputee subjects to marginalise the effect of the users' behaviour on the image acquisition step. One potential influence is the distance between the camera and the object that can be changed by the user during the realtime

Between-Object Cross-Validation

A)

B)

Figure 5. Between-object cross-validation. All objects are in the palmar wrist pronated class and all belong to the Newcastle Grasp Library. In each of the 10 repetitions, out of the 11 available objects, 4 were randomly selected for cross validation. All original images were downsampled before further analysis. (A) Training set. (B) Testing set.

Table 3. The number of objects in each class used as training and testing sets in the BOC analysis.

Grasp type \ Database	ALOI		Newcastle	
	Train	Test	Train	Test
Pinch	81	9	15	4
Tripod	147	16	7	4
Palmar wrist neutral	75	8	26	4
Palmar wrist pronated	124	13	7	4

experiment. Other influences may be participant's motivation, the quality of the EMG signals and physical fatigue.

To perform this test, we used an inexpensive web camera (Logitech Quickcam® Chat), instead of the high-resolution DSLR Canon camera that we used previously to make Newcastle dataset. The webcam was attached to a photography tripod stand. The distance between the webcam and the object was fixed at 60 cm and the webcam was 15 cm higher than the target object such that we could take pictures in the same way as we took in the Newcastle grasp library. The camera was connected to the recording laptop through a USB link. The imaging resolution was set to 640 × 480 pixels.

With clicking on a command button on a MATLAB®-based graphical user interface (GUI), an image was acquired

Image Preprocessing

A)   B)

C)   D)

E)   F)

G)   H)
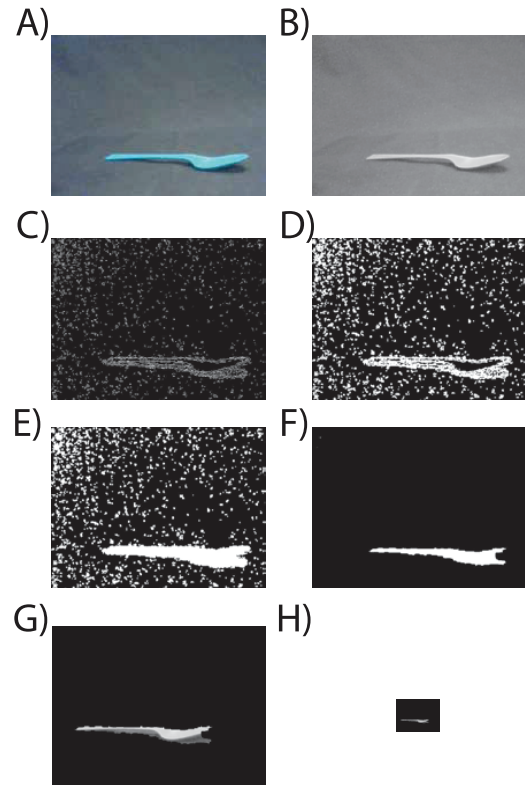
Figure 6. Image preprocessing: (A) original image, taken by the webcam, (B) grey-scale transformation, (C) sobel edge detection, (D) dilation, (E) filling the closed spaces, (F) erosion and filtering the extra noises, (G) multiplication of the mask calculated in F to the original image in A and translation to the lower centre of the image, (H) downsampling to 48 × 36 pixels.

and a series of image processing operations were executed to detect the object in the scene and remove the background. Figure 6 illustrates all of the preprocessing steps. The output of final step, that is G, was resized to 48 × 36 pixels and then normalised according to section 2.2; before feature extraction and classification. Preprocessing was required to remove the background.

We used a two-layer CNN trained for the realtime tests. The test process was repeated for 7 different random views of 24 objects (6 in each grasp group). In this analysis, 16 out of the 24 (66%) objects were not seen by the trained CNN and hence were novel.

All offline and computer-based realtime tests were implemented in MATLAB® in a personal computer with an Intel Core i5-47670 CPU (3.4 GHz), running a 64-bit Windows 7 operating system, with 32 GB RAM.

2.7. Realtime test platform with amputee users in the loop

2.7.1. Subjects. The experiment was conducted with two amputee volunteers who use split hook prostheses in daily life. At the time of this study, their experience of using myoelectric hands was limited to only our laboratory-based experiments. Further information is available in table 4.

Table 4. Amputee volunteers' information.

Identifier	Gender	Age	Cause of amputation	Years since amp.	Missing limb	Prosthesis use
M	Male	28	Car accident	7	Right	Split hook
D	Male	54	Cancer (epithelioid sarcoma)	19	Right	Split hook
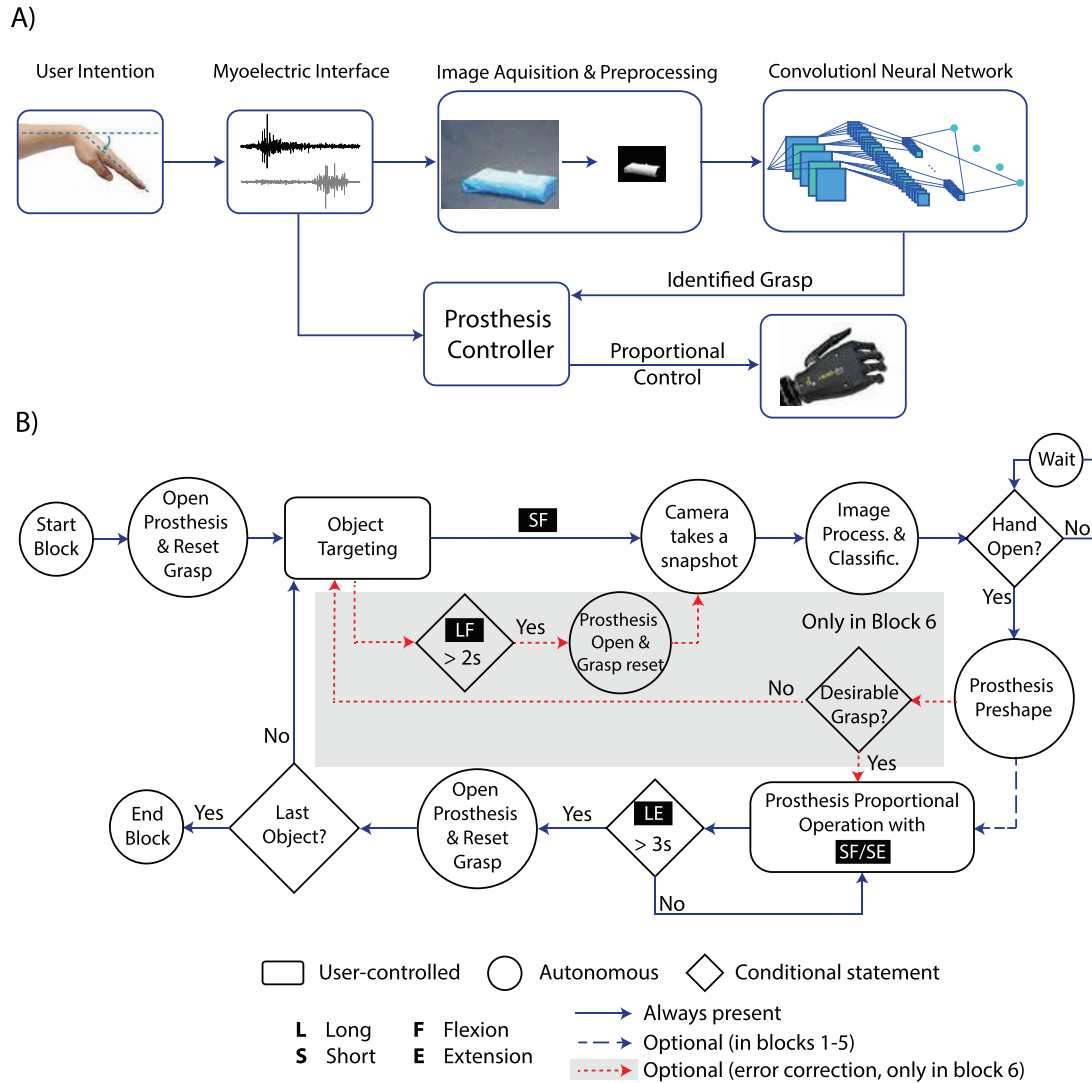
Figure 7. Overall control structure. (A) A block diagram representation of the method; (B) detailed programme flow that was operated via a standard two-channel myoelectric interface.

The study was approved by the Newcastle University ethics committee and carried out at the School of Electrical and Electronic Engineering. Participants signed the experiment consent form.

As in the computer-based realtime experiments, participants sat such that the prosthesis webcam was roughly 60 cm away from and 15 cm higher than the object. Before the start of the experiment, we confirmed that at this distance, they could maintain a comfortable posture to take a picture with the camera and reach the objects readily. In each trial subjects reached one object. As such, in none of the trials the object of interest was occluded by any other objects.

2.7.2. Overall control structure and system components. A general flow digram for our realtime experiment is shown in figure 7(A). Figure 7(B) illustrates the implemented

programme. In the following, we describe the main components of the programme flow. As we will fully describe in section 2.8, the realtime experiment consisted of 6 blocks. The main difference between blocks 1-5 and block 6 was that in the last block an error correction routine was enabled. This additional feature was achieved with the linkages and operations within the grey box in figure 7(B). These connections were inactive in blocks 1–5. Otherwise, all blocks used the same programme for controlling the prosthesis.

A short (300 ms) flexion of wrist muscles was required to trigger the webcam to take a snapshot. After a grasp is identified, the prosthesis was controlled proportionally according to the input EMG signals recorded from the wrist flexor and extensor muscle groups. Long (3 s) extensions reset the grasp and opened the prosthesis.

Table 5. Summary of datasets used in different experimental conditions for training and testing the CNN structure. NCL stands for the Newcastle grasp library. All images in the ALOI + NCL database have been taken with the DSLR camera.

Condition \ Database	Train database	Test database
Offline	ALOI + NCL	ALOI + NCL
Realtime, computer	ALOI + NCL	Webcam
Realtime, amputee	ALOI + NCL	Webcam

In block 6, if the user did not approve the classifier output, they could re-aim the prosthesis at the object and issue a long (2 s) flexion of the wrist muscle to re-open prosthesis, reset the grasp and take a new snapshot. From that point onwards, the control mechanism was exactly as it was in blocks 1–5. The user could repeat this error correction approach until an appropriate grasp is identified.

2.7.3. Myoelectric control. The EMG signals were recorded with two Delsys® Trigno™ lab wireless EMG electrodes. The electrodes were placed on the wrist flexor and extensor muscle groups on the forearm after skin preparation. Surface EMG signals were band-pass filtered between 20 Hz and 450 Hz before sampling at 2 kHz via a Trigno Digital SDK, executed under MATLAB®.

The EMG signals were then transformed into analogue control signals such that 0 and 1 represented the EMG at rest and at a comfortable level of contraction (typically 10–15% of the maximum voluntary contraction) respectively. Generating muscle activity at this low amplitude may be sensitive. However, as we have demonstrated earlier [10, 24, 53–55], with practice participants can learn to contract their muscles reliably at this low level of the MVC to perform a computer task or to control a prosthesis. One reason may be that the magnitude of the signal-dependent motor noise at such low percentages of the MVC is very small [56].

For each EMG channel, a control signal *c* was computed every 100 ms by smoothing (with a rectangular window) the preceding 500 ms of rectified EMG after correction for offset according to

$$c_k = \alpha_k \sum_{-500 \text{ ms}}^{\delta=0} |\text{EMG}_k(t+\delta)| \qquad (8)$$

where $|\text{EMG}_k(t)|$ denotes the rectified activity of muscle *k* at time *t*. The coefficient α_k normalises the control signal by muscle activity at the comfortable contraction level. During a short (15 min) 'familiarisation and calibration' block, subjects were provided with visual feedback of the raw EMG data in two channels and asked to imagine flexion and extension of the wrist alternatively. We ensured that both participants were able to contract the two muscles groups independently before further calibration. To that end, we empirically determined a separate threshold activity for the two control signals. With provision of realtime feedback on a computer screen, we asked the participants to activate one muscle group and cross the corresponding control signal above the threshold whilst keeping the control signal of the other muscle group below its threshold. More details with regards to the calibration can be found in our earlier

work [53, 54]. For subject D, the control signal was recalibrated due to a posture change half-way in the experiment.

2.7.4. The i-limb ultra prosthesis. An open source i-limb Ultra prosthesis (Touch Bionics, an Össur HF company) was used in this work. A MATLAB-based driver was developed that enabled proportional control of individual digits wirelessly via Bluetooth. The hand was powered with a pair of 7.4 V rechargeable batteries.

2.7.5. Wrist rotator. A prosthetic wrist rotator (Motion Control, Inc, USA) was used to enable clockwise and counter-clockwise rotation of the i-limb. The wrist was actuated via an in-house built bidirectional (H-bridge) drive mechanism. The wrist was powered with a doubly insulated power supply set to 7.4 V and rotor direction was controlled via rectangular TTL (5 V) pulses generated with a USB-6002 data acquisition system (National Instruments, USA).

2.7.6. Webcam. The same webcam (Logitech Quickcam® Chat) was used in the computer-based experiment and experiments with amputees. In the latter case, it was attached to the dorsum of the i-limb by means of double-sided velcro. A USB link connected the webcam to the recording laptop. The imaging resolution was set to 640×480 pixels. For analysis, images were downsampled to 48×36 pixels after grey-scale conversion.

2.8. Experimental protocol

The realtime experiment comprised 6 blocks of a pick and place task. In each block, subjects grasped, moved and placed 24 objects. The order of objects in blocks was pseudo-randomised. This order however remained unchanged between blocks and subjects. In each trial the experimenter placed the object at the standard distance on the table in front of the participant.

In blocks 1 and 2, subjects had realtime visual feedback of the measured raw EMGs as well as the calculated control signals on a computer screen. In addition, they could see the webcam video stream, the snapshot that they took and the classification outcome. In blocks 3 and 4, only the raw EMG signals and the control signals were presented as feedback. In block 5, subjects had no computer-based visual feedback at all. Finally, in block 6, similar to block 5, subjects had no visual feedback. They however could reject the grasp identified by the classifier by re-aiming the webcam at the object to take a new picture. This allowed the CNN structure to classify the new image and identify the correct grasp. Due to technical reasons, subject D could not use the error correction function.

With this arrangement of blocks, we combined the familiarisation and testing steps such that the experiment was as short as possible. We therefore analysed the data from familiarisation blocks 1 to 4 as well as data in blocks 5 and 6.

For the experiment with subject M, we allocated a fixed 3 s interval in the beginning of each trial to provide enough time for the participant to settle into the trial before activating the muscles. After the first few trials, we realized that this indeed was a sub-optimal approach because the subject enthusiastically flexed the
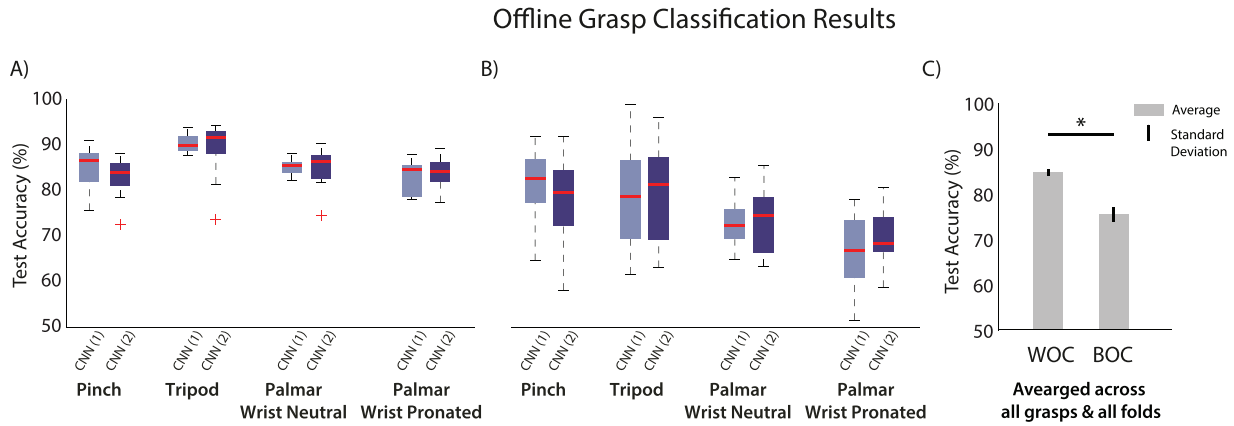
Figure 8. Offline experiment decoding performance comparison. A and B: balanced classification accuracies for within- (left) and between- (middle) object cross-validations (10 folds). CNN(1) and CNN(2) represent one-and two-layer CNN structures, respectively. Boxplot description: horizontal red lines, medians; solid boxes, interquartile ranges; whiskers, overall ranges of non-outlier data; red crosses (+), outliers. C: summary of the within- and between-object cross-validations performance in terms of average classification accuracy together with standard deviations. ∗ denotes statistical significance. (A) Within Object Cross-validation (WOC). (B) Between Object Cross-validation (BOC). (C) WOC versus BOC (summary).

wrist flexor muscles very early to take a picture, before the end of this period. In the experiment with subject D, the protocol was changed slightly such that an audio beep cued the start of the trial, instructing the subject to flex the wrist flexor muscles to activate the webcam. In addition, we made the prosthesis preshaping period shorter to improve responsiveness. As it will be seen in the result, the choice of the trial start protocol and preshaping time affected the total trial duration. However they did not influence the answer to the main question of this work, that is, whether the deep learning structures can be used to offer grasp classification without explicitly measuring object dimensions.

The realtime test was implemented in MATLAB® on a Lenovo laptop with an Intel Core i7-4559U CPU (2.10 GHz), running a 64-bit Windows 7 operating system, with 8GB RAM.

Table 5 summarises the datasets used in different experimental conditions for training and testing the CNN structure.

3. Results

In this section, three categories of results are presented. The first set of results are offline grasp classification scores. For this analysis we used the images of the ALOI database together with the high-resolution images collected for the Newcastle Grasp Library. The aim of this analysis was to test the idea of grasp identification with CNN, fine-tune the CNN structure and identify the most effective classification architecture for the realtime experiments. The second set includes the classification results of the computer-based realtime experiments in which all images were taken with the webcam. The third set of results reports the performance achieved by the amputees in using the proposed deep learning-based vision system for prosthetic grasp in the realtime scenario.

### 3.1. Offline grasp classification

Figure 8 shows the results of the WOC and BOC cross-validation schemes. Both were performed on the combined ALOI and Newcastle libraries. We compared the results of the one-layer and two-layer CNN structures.

A repeated measure two-way ANOVA test revealed no statistical difference between the classification scores for the results achieved by using a one- (80.0%) or a two-layer (79.9%) CNN feature extraction structures ($n = 10$, $F_{1,9} = 0.001$, $p = 0.98$). Figure 8(C) however shows that the difference between the average classification scores for the main effect of the cross-validation type (WOC: 85.29% versus BOC: 74.74%) was statistically significant ($n = 10$, $F_{1,9} = 32.08$, $p < 10^{-3}$). This was predictable since generalisation across views of an object would be less challenging than generalisation to novel objects in the BOC case.

Specifically, in the BOC setting, the two-layer CNN structure led to 0.7% (1-layer: 74.38%, 2-layer: 75.10%) higher classification score when compared to the one-layer CNN setting. This difference was not statistically significant (post-hoc analysis with a paired t-test, $t_9 = 0.28$, $p = 0.78$). For the following realtime experiments, we chose to proceed with the two-layer CNN setting, due to better average performance in three of four grasp classes, figure 8(B).

3.2. Computer-based realtime performance analysis

Figure 9 demonstrates the classification performance achieved in a realtime, but computer-based, setting. For this realtime experiments, one of the ten aforementioned trained CNN structures, that presented a reasonable grasp classification of novel objects during offline BOC tests, was selected. As such, we adopted the CNN parameters that resulted in an average performance of ∼70%; from within a range of settings that gave performances between 64% and 75%. Having six distinct objects in each grasp group and examining seven *random* views of each enabled us to simulate a real scenario closely before bringing the variability caused by the user into account. In figure 9, the proposed grasp for each object and view is shown. In an ideal case, that is 100% correct grasp classification, each bar would be in a single colour. Emergence of different colours indicates incorrect classification.

Real-time Grasp Classification Results (Computer-Based)
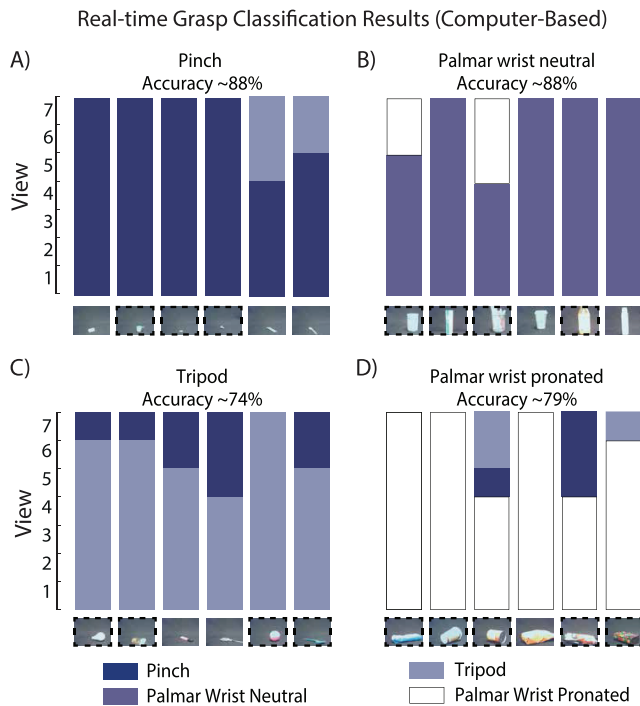
Figure 9. Two-layer CNN architecture average classification performance for four grasp types in on-line computer-based test. All images were converted to grey-scale and downsampled before further analysis. Objects shown with dashed black box around them were novel to the classifier. All other objects were seen by the classifier however they were rotated randomly for this test. In the case of 100% correct classification, each bar would be in a single colour.

With this computer-based test, we quantified the time taken for the system to identify a grasp (correct or incorrect) from a low-resolution input image. The time periods needed for feature extraction with the CNN structure and classification were 78 ± 6 ms and 3 ± 0.03 ms, respectively.

3.3. Realtime test platform with an amputee user in the loop

We tested the whole system with two trans-radial amputee volunteers. Figure 10 shows few representative trials including the recorded myoelectric signals, the acquired images and classification results. This data is from the experiment with subject M. Figure 10(A) illustrates a trial in which the participant oriented the hand such that a reasonable image of the object was acquired; image pre-processing and grasp classification were performed accurately and the correct grasp was identified. In this trial, the subject exhibited an average performance in the pick and place operation ($\sim$7 s). Figure 10(B) shows a trial in which an incorrect classification took place, that is a palmar wrist pronated instead of a tripod. The participant however accepted the incorrect grasp and accomplished the trial. Figure 10(C) shows an example of the trial that the classification was incorrect initially, because the hand was not oriented in a way that the object was fully in the scene. Repeated efforts by the participants were unsuccessful until the fourth time the participant took a snapshot. Once the correct grasp, that is palmar wrist neutral, was selected, the participant completed the trial.
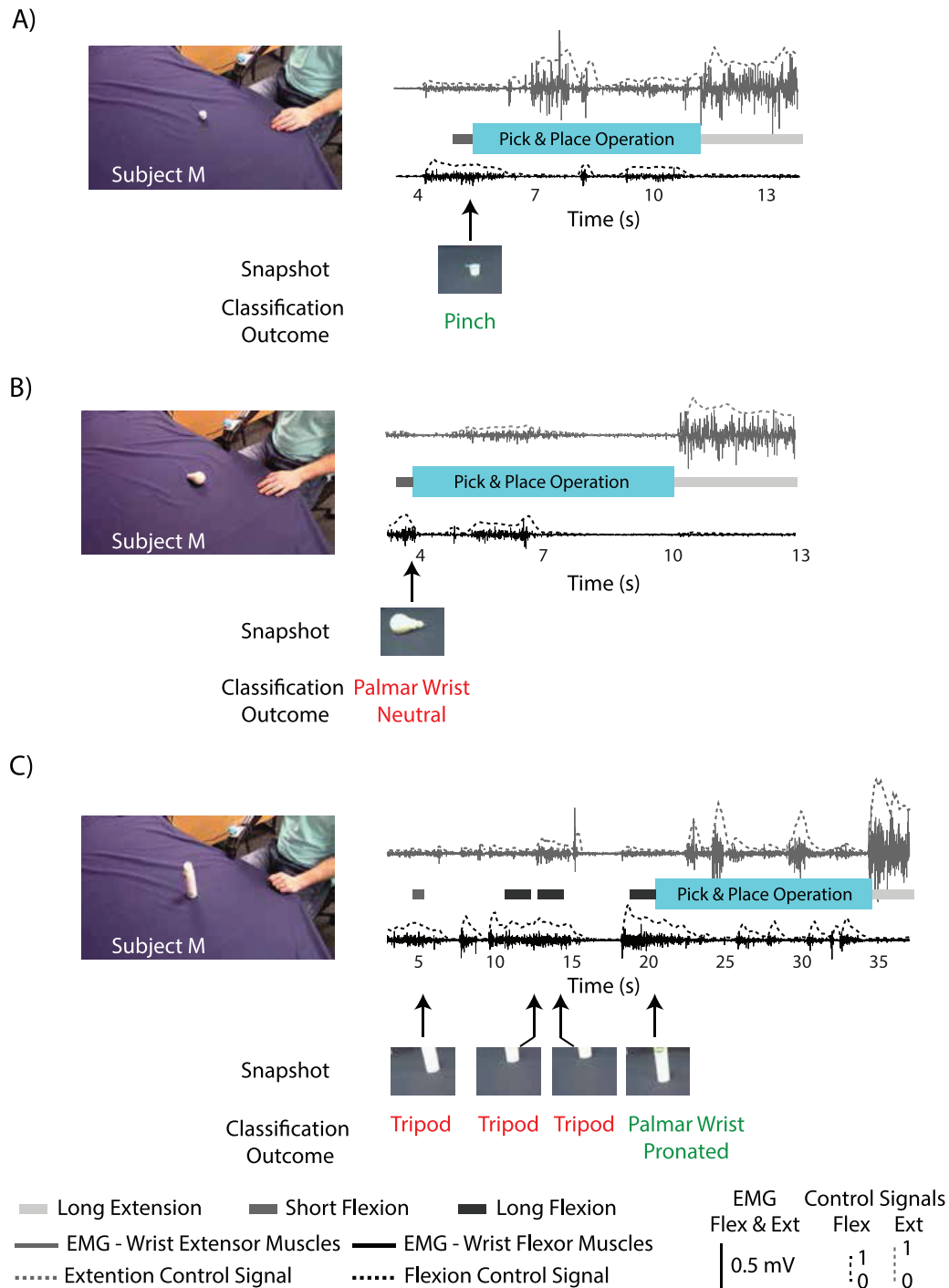
In the realtime experiments with amputee subjects, we included 8 seen, but randomly-rotated, objects as well as 16 novel objects. With this setting, we tested in realtime both within- and between-object generalisation. Figure 11 illustrates a summary of all results in the realtime experiment for the two volunteers M (left column) and D (right column). Figure 11(A) shows the classification accuracy achieved in each block with respect to individual classes. Importantly, in blocks 1 to 5, we report the percentages of correct classification, that is we only consider trials for which the identified grasps matched exactly with the labels that we assigned to that particular object. For block 6, the same terms apply except that the classification results are reported for the final attempt that the user made in each trial; as error correction was enabled. Figure 11(B) shows the overall accuracy in blocks 1 to 6 and in block 6 only, across all grasps. In addition, we have reported the percentage of trials in which the classification was incorrect, however, the subjects accepted the offered grasp and finished the trial successfully or did not accept the offered grasp. In the latter case, if the subject could not complete the trial, the experimenter stopped the trial. Participants were on average more successful in block 6 when compared to the average performance in all blocks: 79% versus 73% for subject M and 86% versus 73% for subject D. When acceptable errors (error subtype 1; as explained in figure 11(B)) included, subject M and D could accomplish 88% and 87% of all trials over the 6 blocks.

Figure 11(C) shows the average trial accomplishment time for each block for participants M and D. For both subjects, block 1 was the longest trial. For subject M, the reduction in the accomplishment time (across the 24 trials) in block 6 versus block 1 was only marginally significant (block 1: 21.4 ± 8.1 s, block 6: 16.7 ± 9.3 s, paired t-test, $n = 24$, $t_{23} = 1.81$, $p = 0.08$). This reduction for subject D, however, was statistically significant (block 1: 30.7 ± 17.2 s, block 6: 19.3 ± 25.7 s, paired t-test, $n = 24$, $t_{23} = 2.26$, $p = 0.03$). This reduction in the accomplishment time was despite the increasing difficulty of the task. As mentioned before, in blocks 3 and 4, the webcam output was not shown on the screen and in the blocks 5 and 6, visual feedback on the screen was withheld totally.

We quantified the time taken for the system to identify a grasp (correct or incorrect) from a low-resolution input image in realtime within our graphical user interface. With the laptop that was used in the realtime experiments the average time needed for pre-processing and classification were 110ms and 40ms, respectively. As mentioned in the Methods section, to take a picture with the camera, subjects had to make a short flexion above the felxion threshold for 300ms, whilst the activity of the extensor muscle group remained below its threshold. As such a correct classification could be achieved within 450ms. All time stamps are shown in figure 11(D).

Finally, we assessed the ability of the proposed structure in generalising to novel objects during the realtime experiments. To that end, for each volunteer we split the results of the realtime experiment for seen and unseen objects in table 6. Out of the 24 objects in each block, eight were seen and 16 were unseen by the trained two-layer CNN. Results showed that it

Figure 10. Three sample trials recorded in the realtime experiments with subject M; A) an example of a successful trial in which the grasp is detected correctly; B) an example trial in which despite the inaccurate classification (palmar wrist pronated instead of tripod grasp) the subject successfully finishes the trial; C) an example of a trial in which the classification is erroneous initially, however, the subject repeats the image acquisition procedure until the correct grasp is identified. (A) A successful trial (Block 6). (B) A trial with an acceptable error in classification (Block 6). (C) A trial with the need to correct the error in image acquisition (Block 6).

was not possible to predict whether classification would be more successful for seen or unseen objects.

4. Concluding remarks

We augmented a commercial prosthetic hand with a webcam and a deep learning-based structure to improve the grasp ability of the amputees. This setting was examined with two trans-radial amputee participants after a comprehensive series of offline and realtime, but computer-based, experiments. We showed that after about an hour of practice, the participant could accomplish 88% of trials successfully.

In current commercial prosthetic hands to switch between the grasp types, the user has to either learn various co-contractions, move the prosthetic hand in certain trajectories or have objects in their environment labelled with RFID tags.
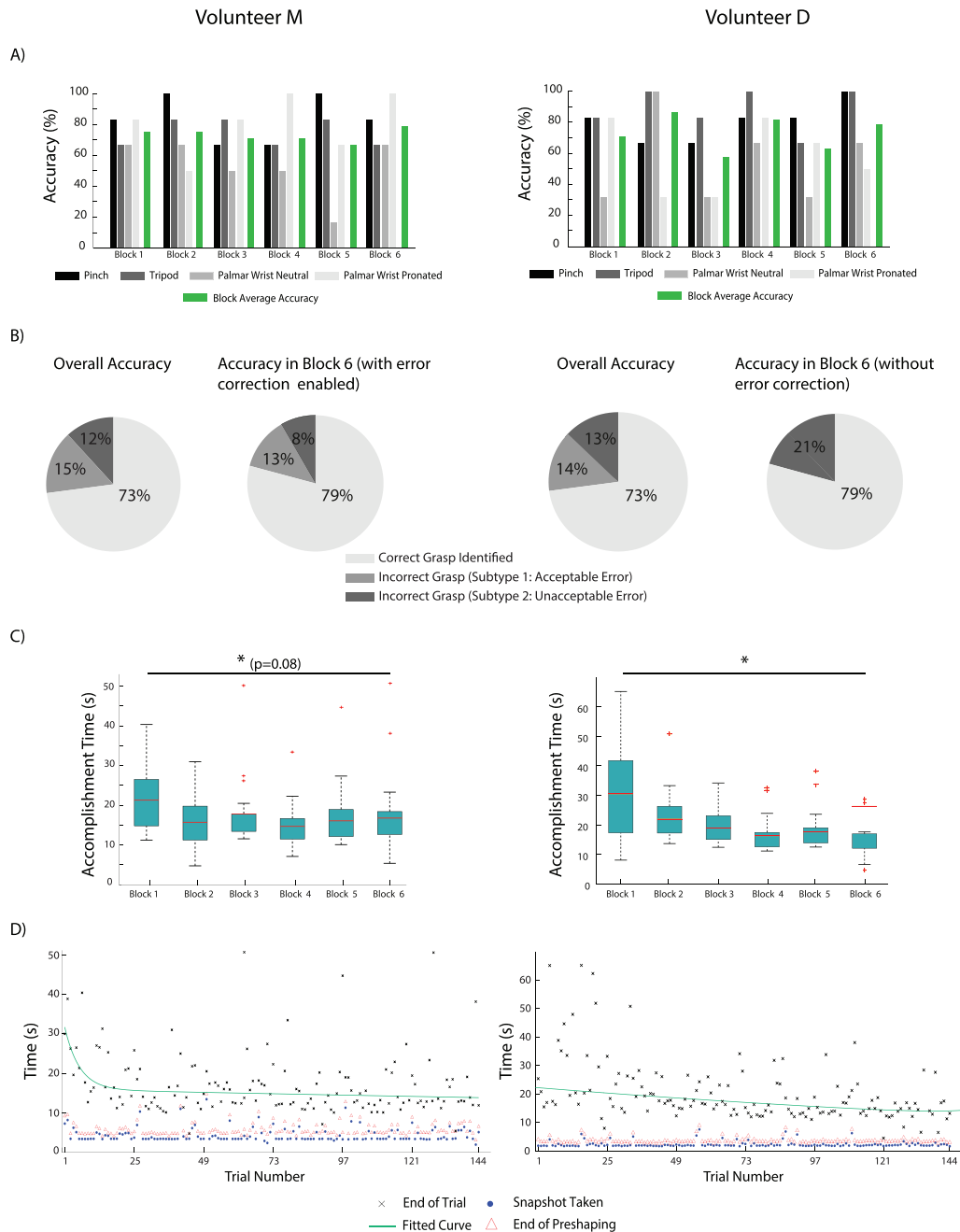
Figure 11. Realtime performance of the proposed system for volunteer M, on the left, and volunteer D, on the right: (A) grasp recognition accuracy performance of each grasp type per block and the overall performance of each block. B) Overall accuracy of the grasp task considering the error being acceptable or not: error subtypes 1 and 2, respectively, and the overall accuracy of in block 6. C) Task accomplishment time comparison between blocks 1 to 6 shown in standard boxplots; D) total trial times with details of the snapshot, the preshape and the end of trial times. ∗ denotes statistical significance. (A) Performance across blocks. (B) Accuracy and error subtypes. (C) Task accomplishment time. (D) Trial times.

These *workaround* techniques have emerged mainly because the promised EMG pattern recognition-based methods have not proved robust, or even feasible, for grasp classification clinically. The non-intuitiveness and shortcomings of the aforementioned approaches have encouraged the emergence of techniques that advocate utilisation of sensing modalities other than the conventional EMG signals, such as accelerometry or in general inertial measurements [14, 25, 26, 57], RFID tags [28], artificial vision including standard cameras as well as

Kinect [29–34]. In almost all multi-modal approaches to control limb prosthesis, it is argued that the incorporation of two or more sources of information can reduce the users' cognitive burden and enhance functionality in terms of accuracy.

In this work, the user could effectively pick objects with 4 different grasp types by capturing a single picture of the object of interest. We adapted and trained a standard CNN architecture to extract abstract grasp-related features of a single low-resolution input object image in realtime.

Table 6. The success rate of each volunteer in the realtime experiments with respect to the objects being seen or unseen. For subject M in Block 6, in which the error-correction was on, we report the performance with respect to the *first* identified grasp*, that is before error correction.

Block \ Volunteer	M		D	
	Seen (%)	Unseen (%)	Seen (%)	Unseen (%)
1	75	75	50	81.2
2	75	75	75	93.7
3	62.5	75	50	56.2
4	37.5	87.5	87.5	81.2
5	75	62.5	62.5	62.5
6	63	75	87.5	75

4.1. Database

In order to train the CNN structure for grasp recognition, a database including a large number of object images was required. Identifying a database with an ample number of household objects can be considered a challenge since most of the accessible databases, e.g. Imagenet [40], include a large variety of objects of which many are not graspable.

In our previous pilot experiment [34], we used the COIL100 database [58] which includes 100 categories of graspable objects. The overall classification performance in the WOC test was 97%. For the BOC test the classification accuracy was 55% with the lowest results in the 'palmar wrist pronated' group. We attributed this poor BOC performance to lacking sufficient number of training objects since other grasp classes that had enough number of training objects gained significantly higher accuracies. Therefore, we used the ALOI database [46] instead, which provided us with more training data in the range of 1000 objects of which we selected 500 objects for analysis. Due to the variety in the number of objects in each grasp group and of course to enable realtime testing where original objects are not available, a database of 71 objects was collected at Newcastle University. These 71 objects were distributed between grasp groups such that they are each provided with sufficient samples for training. This augmented image database was used for CNN training in this work.

4.2. Object classification versus grasp identification

We adapted the CNN architecture for grasp recognition rather than object identification. Supervised learning systems (e.g. [59, 60]) including the CNN setting (e.g. [61, 62]), lack the capability of generalisation to novel objects, which is a crucial requirement in prosthetic hand applications. To address this issue, we either need a very large amount of training data or we can capitalise on the flexibility of deep learning system to generalise based on learning abstract representation of different classes of training data. Forming large image libraries can be challenging since it requires advanced hardware for photography and computing facilities for data handling and storage. Instead, we approached the problem by noting that rather than having a large number of classes of objects, we can group the objects according to their most appropriate grasp type. In this way, the output space includes only a small number of grasps. Consequently, the detection task can be generalised to unknown objects and any type of objects can be detected and classified correctly.

4.3. The CNN design considerations

The difference between the two CNN structures, that is 1- and 2-layer, was very small. Despite no statistical difference in offline analysis between the two cases we chose to use a 2-layer structure as it showed a slightly better performance in three of four grasp classes (figure 8(B)). Realtime implementation on a laptop was similar with differences in the nanoseconds range. In principle, with using a smaller network one could avoid over-fitting. We avoided over-fitting in the 2-layer network by using Tikhonov regularisation in training the CNN structure where the matrix $\mathbf{W}_j^l$ in the last layer is optimised.

Tuning and training of a CNN structure may be very time-consuming. However once trained, it offers a very fast response time. Typical times for training the proposed 2-layer CNN structure were about 2 hours without a GPU. In fact, the slowest component of the proposed approach is the image pre-processing block that takes ∼110ms to carry out all steps that were introduced in figure 6. For the realtime experiments, we used standard MATLAB instructions without additional GPU hardware. With the advent of fast GPU chips, that to the mobile phones industry, we believe that realtime implementation of our standard image preprocessing tasks will be much faster.

Although not included in the results section, we tested the hypothesis that using a pre-trained CNN, for example with all images in the ImageNet database [40] could enhance the classification accuracy. We therefore re-tuned a ResNet18 [61], an 18-layer network pre-trained with ImageNet, with the combined ALOI and Newcastle images and then repeated the BOC test. We observed a large reduction in the classification scores to ∼50%. Such a poor performance may be because many of the objects in the ImageNet database are not graspable, e.g. an airplane or a tree. In addition, most images include significant clutter and have various backgrounds, e.g. an ambulance on a street. Whilst these results are not in favour of using a pre-trained network, we do not rule out the possibility that pre-trained architectures can be used to enhance the generalisation performance. Perhaps the use of pre-trained networks, that are trained with a large number of graspable objects can lead to higher performance.

We sought to understand whether object specific patterns were extracted by the CNN structure for grasp classification or the size and orientation of the objects enable the CNN setting to generalise. Figure 12 illustrates two examples per grasp class. The 25 maps per object are the outcome of the second convolution layer of the CNN architecture after the ReLU stage, as introduced in figure 3. This preliminary visualisation suggests that the determining factors for classification and generalisation are the size and the orientation of the object. Further work may need be to verify the consistency of this finding in a larger number of objects and object views.
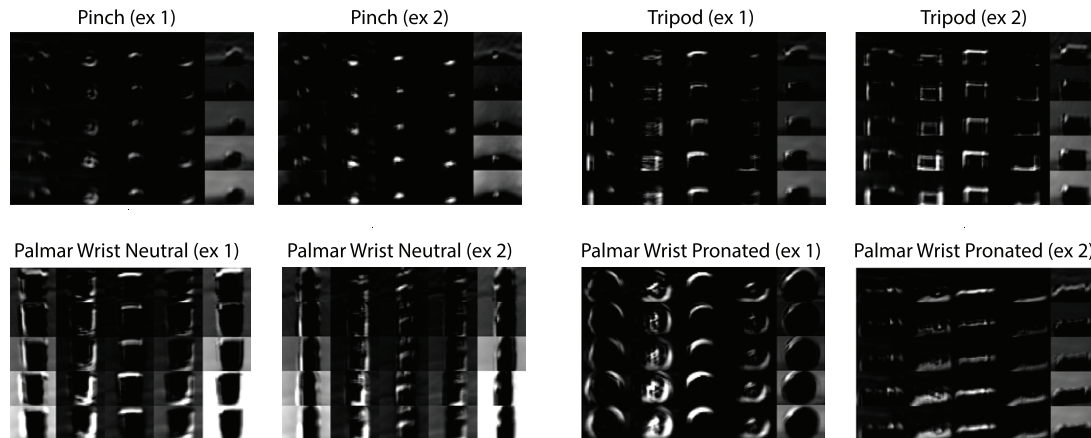
Figure 12. Two examples for each grasp class. After convolving to the second filter each resultant image, is transformed into 25 feature maps whilst passing through ReLU and before being sub-sampled. These maps suggest that generalisation may be achieved because the abstract object features are size and orientation of the objects.

4.4. Classification and an alternative approach for error correction

The CNN architecture can have several layers of convolution and pooling. The last layer should be fully connected, e.g. a neural network. We used the Softmax regression classifier. The integrated use of the CNN and Softmax systems is in line with the conventional approaches in the deep learning community [40, 42, 63]. Instead of the Softmax classifier, other classifiers could have been used. Comparing the performance of different classifiers falls outside the remit of this work.

The output of the Softmax classifier provides the probability for each class. Therefore, when the most probable grasp is not suitable, other grasp types of decreasing probability may be provided. This can feature as an automatic error correction mechanism. However, there were two main reasons behind our decision for not using this approach though we find it very interesting and feasible from an engineering point of view. These reasons include:

- The addition of an artificial vision system for artificial hands makes the prosthesis more autonomous [31] and less under the control of the subject. Our initial and unbiased briefing of the subjects with regards to the experiment suggested that they both would like to have a degree of control over the function of the prosthesis. As such, we decided to test the manual error-correction method only. In this approach, in block 6, subjects could restart the process by resetting the prosthesis and taking a new snapshot.
- Our volunteers were both naíve to the concept of the experiment and neither used myoelectric prosthesis in daily life. In addition, our experiment was already rather long (+2 hours). Therefore, the addition of another condition to the experiment, in which errors are automatically corrected, would tire the participants.

4.5. Feasibility of more grasp types

We limited the number of grasp types to four. The number of grasps however can be increased readily upon availability of training data. For instance, lateral grasp is not included

currently within the grasp types. Objects that are grasped with a lateral grip, e.g. a credit card, present typically a distinctive flat shape which is different from our training images. Therefore, with augmenting the existing database with images of objects requiring a lateral grasp, we can include the lateral grasp as an additional grasp class. Whether prosthesis users would use more than four or five grasps will remain to be investigated.

4.6. Performance in the presence of clutter

Identification and segmentation of an object in a cluttered scene or when the object lies on an arbitrary background can be an extremely challenging computer vision task. In this proof-of-principle work, we tested the use of deep learning algorithm in a clutter-free environment. Previous work such as in [30, 31] incorporated 3D point clouds to segment the scene (in addition to colours and edges) to facilitate segmentation. One interesting study would be to combine the two approaches and use the 3D features as inputs into the CNN system.

4.7. Realtime performance: computer-based versus human experiments

With the computer-based realtime experiments, we simulated a grasp classification scenario without having the user in the loop. We believe that it was an appropriate practice since it gave us an indication of realtime performance without biases induced by the user, e.g. camera view and distance to the object. The computer-based results with the average performance of 84% were higher than the accuracy achieved in the realtime experiments with amputee subjects in the loop specially in early blocks. With training, both subjects improved performance yet they fall short of the score that was achieved in the computer-based experiment. We believe that the higher performance in the computer-based experiment was because the camera view and distance to the objects were fixed during testing. Other intrinsic factors, such as physical and mental fatigue, can deteriorate realtime performance. Further

investigations are needed to identify underlying sources of error and inaccuracies during realtime experiments; be it in the laboratory, clinical or real-life settings.

In the realtime experiments with amputee subjects, out of the 24 objects in each block, eight were seen and 16 were unseen by the trained two-layer CNN. Results did not show that generalisation to unseen objects was necessarily less successful than classification of seen objects. This is in contrast to what we observed in the offline experiments in which the BOC performance was lower than that of the WOC performance. However, this finding corroborates earlier work in [29–31] supporting the hypothesis that users' behaviour could play an important role in the accuracy of vision-based prostheses control architectures.

4.8. User training with full or partial visual feedback

The webcam was mounted on the dorsum of the i-limb hand. This was in line with earlier work on vision-based prosthesis control [29]. However, in more recent work, Marković *et al* [31] placed the sensors on the user to facilitate targeting the object. We showed that with training in a step-by-step approach (blocks 1-6) subjects can learn to target the object accurately such that all of the object boundaries are in the scene. This is particularly important for tall objects in the palmar wrist neutral group (figure 10(C)). Following the familiarisation block, and the first two measurement blocks, the visual feedback from the webcam output was withheld, however the performance did not drop. The available proprioceptive feedback from the arm and the truck muscles may have facilitated accurate targeting.

4.9. User feedback

Both subjects provided positive feedback on the use of the proposed vision-enabled prosthetic control approach. For instance, subject D said: '*Just getting the routine was difficult at the beginning but once this was established it became much easier. If it would be further refined (in terms of positioning of camera) I would certainly use this and always give feedback*'. Subject M tested the proposed approach and a novel pattern recognition system on the same day. When asked which of the two approaches he would prefer, he replied: '*I'd like the pattern recognition better, when it works perfectly! For the time being, the vision-based system seems to be a good solution. I liked its responsiveness very much*'.

4.10. Directions for further development

In the proposed setting, misclassification could stem from inaccurate object detection or from insufficient feature extraction by the CNN structure. Advanced image processing techniques may be used to address the former. The latter problem may be dealt with fine-tuning the CNN parameters according to an objective criterion. Beyond these challenges, one critical issue that can affect the performance of *any* vision-based prosthetic control system is the distance between the object of interest and the camera. Previous work incorporated additional sensors, e.g. sonar [29] or stereovision [31], to alleviate the uncertainty with regards to the true object sizes. Our current work includes using movement inertial measurements during reach to approximate the distance from the target object and rescale the images before giving them to the CNN structure.

Acknowledgments

References

[1] Twiste M 2011 Limbless Statistics, United National Institute for Prosthetics & Orthotics Development

[2] Ziegler-Graham K, MacKenzie E J, Ephraim P L, Travison T G and Brookmeyer R 2008 Estimating the prevalence of limb loss in the United States: 2005 to 2050 *Arch. Phys. Med. Rehabil.* **89** 422–9

[3] Oskoei M A and Hu H 2007 Myoelectric control systems—a survey *Biomed. Signal Process. Control* **2** 275–94

[4] Farina D, Jiang N, Rehbaum H, Holobar A, Graimann B, Dietl H and Aszmann O 2014 The extraction of neural information from the surface EMG for the control of upper-limb prostheses: emerging avenues and challenges *IEEE Trans. Neural Syst. Rehabil. Eng.* **22** 798–809

[5] Nazarpour K, Cipriani C, Farina D and Kuiken T 2014 Advances in control of multi-functional powered upper-limb prostheses *IEEE Trans. Neural Syst. Rehabil. Eng.* **22** 711–15

[6] Gonzalez-Vargas J, Dosen S, Amsuess S, Yu W and Farina D 2015 Human-machine interface for the control of multi-function systems based on electrocutaneous menu: application to multi-grasp prosthetic hands *PLoS One* **10** e0127528

[7] Englehart K and Hudgins B 2003 A robust, real-time control scheme for multifunction myoelectric control *IEEE Trans. Biomed. Eng.* **50** 848–54

[8] Hargrove L J, Englehart K and Hudgins B 2007 A comparison of surface and intramuscular myoelectric signal classification *IEEE Trans. Biomed. Eng.* **54** 847–53

[9] Nazarpour K, Sharafat A R and Firoozabadi S M P 2007 Application of higher order statistics to surface electromyogram signal classification *IEEE Trans. Biomed. Eng.* **54** 1762–69

[10] Pistohl T, Cipriani C, Jackson A and Nazarpour K 2013 Abstract and proportional myoelectric control for multi-fingered hand prostheses *Ann. Biomed. Eng.* **41** 2687–98

[11] He J, Zhang D, Jiang N, Sheng X, Farina D and Zhu X 2015 User adaptation in long-term, open-loop myoelectric training: implications for EMG pattern recognition in prosthesis control *J. Neural Eng.* **12** 046005

[12] Smith L H, Kuiken T A and Hargrove L J 2015 Use of probabilistic weights to enhance linear regression myoelectric control *J. Neural Eng.* **12** 066030

[13] Kamavuako E N, Scheme E J and Englehart K B 2016 Determination of optimum threshold values for EMG time

domain features; a multi-dataset investigation *J. Neural Eng.* **13** 046011

[14] Khushaba R N, Al-Timemy A H, Kodagoda S and Nazarpour K 2016 Combined influence of forearm orientation and muscular contraction on EMG pattern recognition *Expert Syst. Appl.* **61** 154–61

[15] Martelloni C, Carpaneto J and Micera S 2015 Classification of upper arm EMG signals during object-specific grasp *Annual Int. IEEE EMBS (EMBC)* pp 5061–64

[16] Castellini C and van der Smagt P 2009 Surface EMG in advanced hand prosthetics *Biol. Cybern.* **100** 35–47

[17] Kakoty N M, Kaiborta M and Hazarika S M 2012 Electromyographic grasp recognition for a five fingered robotic hand *Int. J. Robot. Autom.* **2** 1–10

[18] Tenore F V G, Ramos A, Fahmy A, Acharya S, Etienne-Cummings R and Thakor N V 2009 Decoding of individuated finger movements using surface electromyography *IEEE Trans. Biomed. Eng.* **56** 1427–34

[19] Al-Timemy A H, Bugmann G, Escudero J and Outram N 2013 Classification of finger movements for the dexterous hand prosthesis control with surface electromyography *IEEE J. Biomed. Health Inf.* **17** 608–18

[20] Morel P *et al* 2016 Long-term decoding of movement force and direction with a wireless myoelectric implant *J. Neural Eng.* **13** 016002

[21] Sears H H and Shaperman J 1991 Proportional myoelectric hand control: an evaluation. *Am. J. Phys. Med. Rehabil.* **70** 20–8

[22] Light C M, Chappell P H, Hudgins B and Engelhart K 2002 Intelligent multifunction myoelectric control of hand prostheses *J. Med. Eng. Technol.* **6** 139–46

[23] Ninu A, Došen S, Muceli S, Rattay F, Dietl H and Farina D 2014 Closed-loop control of grasping with a myoelectric hand prosthesis: which are the relevant feedback variables for force control? *IEEE Trans. Neural Syst. Rehabil. Eng.* **22** 1041–52

[24] Pistohl T, Joshi D, Ganesh G, Jackson A and Nazarpour K 2015 Artificial proprioceptive feedback for myoelectric control *IEEE Trans. Neural Syst. Rehabil. Eng.* **23** 498–507

[25] Atzori M, Gijsberts A, Castellini C, Caputo B, Hager A-G M, Elsig S, Giatsidis G, Bassetto F and Müller H 2014 Electromyography data for non-invasive naturally-controlled robotic hand prostheses *Sci. Data* **1** 140053

[26] Krasoulis A, Vijayakumar S and Nazarpour K 2015 Evaluation of regression comparison for the continuous decoding of finger movement from surface EMG, accelerometry *IEEE Neural Engineering and Int. Conf.* pp 631–4

[27] Cho E, Chen R, Merhi L-K, Xiao Z, Pousett B and Menon C 2016 Force myography to control robotic upper extremity prostheses: a feasibility study *Frontiers Bioeng. Biotechnol.* **4** 18

[28] Trachtenberg M S, Singhal G, Kaliki R, Smith R J and Thakor N V 2011 Radio frequency identification—an innovative solution to guide dexterous prosthetic hand *Annual Int. IEEE EMBS (EMBC)* pp 3511–14

[29] Došen S and Popović D B 2011 Transradial prosthesis: artificial vision for control of prehension *Artif. Organs* **35** 37–48

[30] Došen S, Cipriani C, Kostić M, Controzzi M, Carrozza M C and Popović D B 2010 Cognitive vision system for control of dexterous prosthetic hands: experimental evaluation *J. Neuroeng. Rehabil.* **7** 42

[31] Marković M, Došen S, Cipriani C, Popović D B and Farina D 2014 Stereovision and augmented reality for closed-loop control of grasping in hand prostheses *J. Neural Eng.* **11** 046001

[32] Krausz N E, Lenzi T and Hargrove L J 2015 Depth sensing for improved control of lower limb prostheses *IEEE Trans. Biomed. Eng.* **62** 2576–87

[33] Marković M, Došen S, Popović D B, Graimann B and Farina D 2015 Sensor fusion and computer vision for context-aware control of a multi degree-of-freedom prosthesis *J. Neural Eng.* **12** 066022

[34] Ghazaei G, Alameer A, Degenaar P, Morgan G and Nazarpour K 2015 An exploratory study on the use of convolutional neural networks for object grasp classification *The 2nd IET Int. Conf. on Intelligent Signal Processing* p 5

[35] Saxena A, Driemeyer J and Ng A 2008 Robotic grasping of novel objects using vision *Int. J. Robot. Res.* **27** 157–73

[36] Kootstra G, Popović M, Jørgensen J A, Kuklinski K, Miatliuk K, Kragic D and Krüger N 2012 Enabling grasping of unknown objects through a synergistic use of edge and surface information *Int. J. Robot. Res.* **31** 1190–213

[37] Lenz I, Lee H and Saxena A 2015 Deep learning for detecting robotic grasps *Int. J. Robot. Res.* **34** 705–24

[38] Kopicki M, Detry R, Adjigble M, Stolkin R, Leonardis A and Wyatt J L 2015 One-shot learning and generation of dexterous grasps for novel objects *Int. J. Robot. Res.* **35** 959–76

[39] Lecun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition *Proc. IEEE* **86** 2278–324

[40] Krizhevsky A, Sutskever I and Hinton G E 2012 ImageNet classification with deep convolutional neural networks *Adv. Neural Inf. Process. Syst.* **25** 1097–105

[41] Razavian A S, Azizpour H, Sullivan J and Carlsson S 2014 CNN features off-the-shelf: an astounding baseline for recognition *IEEE Conf. on Computer Vision and Pattern Recognition Workshops* pp 512–19

[42] Deng L 2014 A tutorial survey of architectures, algorithms, and applications for deep learning *APSIPA Trans. Signal Inf. Process.* **3** e2:1–29

[43] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S and Darrell T 2014 Caffe: convolutional architecture for fast feature embedding *Proc. ACM Int. Conf. Multimedia* pp 675–78

[44] Oquab M, Bottouand I L and Sivic J 2014 Learning and transferring mid-level image representations using convolutional neural networks *Proc. IEEE Conf. Computer Vision Pattern Recognition* pp 1717–24

[45] Zeiler M D and Fergus R 2014 Visualizing and understanding convolutional networks *Proc. 13th European Conf. Computer Vision: Part I* pp 818–33

[46] Geusebroek J-M, Burghouts G J and Smeulders A W M 2005 The Amsterdam library of object images *Int. J. Comput. Vis.* **61** 103–12

[47] Nair V and Hinton G 2010 Rectified linear units improve restricted Boltzmann machines *Proc. Int. Conf. Machine Learning* pp 807–14

[48] Ranzato M, Huang F J, Boureau Y and LeCun Y 2007 Unsupervised learning of invariant feature hierarchies with applications to object recognition *Int. Conf. on Computer Vision and Pattern Recognition* pp 1–8

[49] Bishop M C 2006 *Pattern Recognition and Machine Learning* (New York: Springer)

[50] Greene W H 2012 *Econometric Analysis* (Boston, MA: Pearson Education)

[51] Bridle J S 1990 Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters *Advances in Neural Information Processing Systems* 2 ed D S Touretzky (San Mateo, CA: M. Kaufmann) pp 211–17

[52] Qian N 1999 On the momentum term in gradient descent learning algorithms *Neural Netw.* **12** 145–51

[53] Nazarpour K, Barnard A and Jackson A 2012 Flexible cortical control of task-specific muscle synergies *J. Neurosci.* **32** 12349–60

[54] Graziadio S, Nazarpour K, Gretenkord S, Jackson A and Eyre J A 2015 Greater intermanual transfer in the elderly suggests age-related bilateral motor cortex activation is compensatory *J. Mot. Behav.* **47** 47–55

[55] Barnes J, Dyson M and Nazarpour K 2016 Comparison of hand and forearm muscle pairs in controlling of a novel myoelectric interface *IEEE Int. Conf. on Systems, Man, and Cybernetics* pp 2846–49

[56] Harris C M and Wolpert D M 1998 Signal-dependent noise determines motor planning *Nature* **20** 780–84

[57] Kyranou I, Krasoulis A, Erden M S, Nazarpour K and Vijayakumar S 2016 Real-time classification of multimodal sensory data for prosthetic hand control *6th IEEE Int. Conf. on Biomedical Robotics and Biomechanics* pp 536–41

[58] Nayar S, Nene S A and Murase H 1996 *Technical Report* CUCS-006-96, Columbia object image library (COIL 100) Department of Comp. Science, Columbia University

[59] Lowe D G 1999 Object recognition from local scale-invariant features *7th IEEE Int. Conf. on Computer Vision* vol 2 pp 1150–57

[60] Alameer A, Ghazaei G, Degenaar P, Chambers J A and Nazarpour K 2016 Object recognition with an elastic net-regularized hierarchical MAX model of the visual cortex *IEEE Signal Process. Lett.* **23** 1062–6

[61] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Int. Conf. on Computer Vision and Pattern Recognition* pp 770–78

[62] Kheradpisheh S R, Ghodrati M, Ganjtabesh M and Masquelier T 2016 Deep networks resemble human feed-forward vision in invariant object recognition *Sci. Rep.* **6** 32672

[63] Sainath N, Mohamed A, Kingsbury B and Ramabhadran B 2013 Deep convolutional neural networks for LVCSR *IEEE Int. Conf. on Acoustics, Speech and Signal Processing* pp 8614–18