Ultrasound image features of the wrist are linearly related to finger positions

Claudio Castellini and Georg Passig

Institute of Robotics and Mechatronics
DLR - German Aerospace Research Center
Oberpfaffenhofen, Germany
e-mail claudio.castellini@dlr.de

***Abstract*—Ultrasound imaging is a widespread technique to gather live images of the interiors of the human body. It is safe and provides high spatial and temporal resolution. In this paper we show that features extracted from the ultrasound section of the human wrist can be used to fully reconstruct the hand movements, including flexion of all fingers and the rotation of the thumb. Surprisingly, it turns out that there is a clear linear relationship between image features and finger positions. The related matrix can be estimated on a rather small subset of samples, and the reconstruction is quite robust across single- and multi-finger movements. This technique can be used to control advanced mechatronic hands, and it finds its paradigmatic application in the case of hand amputees.**

I. Introduction

Developed soon after the Second World War as a diagnostic device, ultrasound imaging (also known as medical ultrasonography, US from now on) is a totally non-invasive technique to visualise structures inside the human body. The general principle is that of wave reflection/refraction: in the modern ultrasound medical devices, an array of piezoelectric transducers is used to generate a focused wave of ultrasound which penetrates the body part of interest; partial reflection of the wave at the interfaces between tissues with different acoustic impedance (density) is then gathered and converted to a grey-scale 2D image. High-grey-valued "ridges" in the image denote tissue interfaces. Modern US machines can achieve sub-millimeter spatial resolution and/or real-time temporal resolution, penetrating several centimeters below the subject's skin. The technique is totally harmless and it has no known side effects, to the extent that one of ist best known applications is the imaging of the foetus with pre-birth diagnostic purposes.

In rehabilitation robotics, especially in prosthetics, this has an immediate application: to use live US images of the residual limb to control the rehab device. US is since a long time successfully used as a diagnostic tool for hand musculoskeletal disorders (e.g., synovitis and rheumatoid arthritis [1], [2], [3]), so it is likely that US images contain enough information to reconstruct — at least partly — the position, velocity and/or force exerted by the fingers. If this intuition is true, and the technology is advanced enough to make it applicable in practice, then US could be used as a means to control a mechanical hand. Moreover, it might be possible to apply the same technique to amputees, according to the severity of the amputation (and therefore to the required position of the transducer on the subject's forearm) and to the residual muscle activity.

This feeling stems from simple observation of the US imaging as the fingers move. The attached movie "example.avi" was gathered from a healthy subject using a standard portable US machine, the transducer lying against the ventral side of the wrist along the transverse plane (orthogonal to the axis of the forearm, see Figure 1, right panel). Even from such a naïve analysis, a clear correspondance between finger movements and deformation of the images is apparent: flexion of the index and pinkie fingers, for example, results in "holes" opening and closing near the surface — since this is a cross-section of the wrist, we are most likely witnessing the contraction of one of the tendons of the *M. Flexor Digitorum Superficialis*. At the same time it must be noted that the deformations associated to finger movements are diverse and complex: sometimes it is a local rotation, sometimes an enlargement/reduction, and sometimes a combination of them. The motions tend to superimpose to one another, and a contracting muscle will shift what is around it in a rather complicated way. Quite clearly, advanced image processing must be employed to solve this problem.

In this paper we show an initial, very promising result along this line. A human subject, wearing a sensorised dataglove, was instructed to mimick with his right hand the movements performed by an animated human hand model on a computer screen. The movements consisted of repeated flexion of the fingers and adduction of the thumb, either one by one or simultaneous. The choice of these six motions is motivated by the fact that they are enforced by the most advanced hand prosthesis of the world at the time of writing, namely the Vincent Hand (Vincent Systems GmbH, www.handprothese.de/vincent-hand).

At the same time, US images of his wrist would be gathered. Offline, local features were extracted from each frame and synchronised with the finger positions as recorded by the dataglove. Statistical analysis reveals that the features are almost perfectly correlated (in the sense of the standard Pearson correlation) to finger positions; and that the correlation is higher where the sections of anatomically relevant muscles appear; for example, pinkie movement is highly correlated with features extracted near the section of the *F.D.Superficialis*, i.e., from the upper-left corner of the images seen in the movie — where the "hole" grows and shrinks.

All in all, it turns out that there is a straightforward linear relationship between the image features and the finger positions, i.e., that $\mathbf{p} = K\mathbf{v}$, where $\mathbf{p} \in \mathbb{R}^6$ represents the position of the fingers, $\mathbf{v} \in \mathbb{R}^n$ encodes the n visual features extracted from the US frames, and K is a $6 \times n$ matrix, estimated with a simple least-square approach from a subset of the $(\mathbf{p}_i, \mathbf{v}_i)$ pairs gathered during the experiment. The relationship is robust across single- and multi-finger movements; for example, a K estimated from index and middle finger movements only can then successfully be used to predict *simultaneous* movement of the index and middle finger.

A. Related work

As far as we know at the time of writing, the only attempt along these lines of research is [4], where significant differences among optical flow computations for finger flexion movements are reported (but not analysed). Optical flow [5] seems not really the best feature choice here, since it is a derivative operator, hard to compute and prone to accumulating integral errors when applied to position recognition.

II. Experimental setup

A. Data gathering

1) Hand motion: an 18-sensor right-handed Cyberglove (Cyberglove Systems, www.cyberglovesystems.com, see also Figure 1, left panel) is used to gather the finger positions. The Cyberglove is a light fabric, rather elastic glove, onto which 18 strain gauges are sewn; the sewing sheaths are chosen carefully by the manufacturer, so that the gauges exhibit a resistance which is proportionally related to the angles between pairs of hand joints of interest. The device can then return 18 8-bit values, proportional to these angles, for an average resolution of less than one degree, depending on the size of the subject's hand, a careful wearing of the glove and the rotation range of the considered joint. (For practical reasons the subject must wear a cotton glove below the Cyberglove; we verified that that would not limit the precision of the device.)

We hereby consider 6 hand motions, namely flexion/extension of the 5 fingers and thumb adduction/abduction. Thumb flexion/extension is roughly equivalent to thumb rotation, indeed a very important motion, characteristic of the high primates and paramount for most activities of daily living.

The above motions are captured by considering the five metacarpophalangeal glove sensors, placed where the proximal phalanxes of the fingers meet the palm, plus the thumb/index abduction sensor for the thumb abduction/adduction. According to the placement of the sensors on the Cyberglove (see Figure 1, center panel), we choose sensors 16, 12, 8, 4 and 0 for the pinkie, ring, middle, index and thumb flexion/extension, and sensor 3 for the thumb rotation. A careful hardware calibration enables us to obtain a resolution of about 7.5 bits over the considered range, actually way below one degree in all cases. Values are normalised between 0 and 1 so that 0 corresponds to the relaxed stance and 1 to the maximum voluntary contraction for the motion under consideration. The six motion values are streamed to a PC at the maximum rate allowed by the glove's underlying serial port connection, namely 88Hz.

2) Ultrasound imaging: US images are gathered using a pre-owned General Electric Logiq-e portable ultrasound machine (see www.gehealthcare.com/euen/ultrasound/products/portable/logiq-e) equipped with a 12L-RS linear transducer. We employ the ultrasound B-mode (the linear transducer scans a plane across the body section) to produce a view of the interior of the forearm at the height of the wrist, along the transverse plane. More precisely, the probe is located at the distal radioulnar articulation (see en.wikipedia.org/wiki/Distal_radioulnar_articulation), at the level of the *Pronator quadratus*.

After an initial round of examinations, the following settings were chosen: ultrasound frequency of 12MHz, minimal onboard image pre-processing (i.e., noise rejection / edge enhancement), focus point at a depth of about 1.3cm, and minimum depth of field ("focus number" set at 1). This results in a frame rate of 28Hz. Since this US machine is not able to stream images over to the PC we employ a VGA frame grabber to grab the US frames across a peer-to-peer Ethernet connection. More details about the image processing appear in the following Section.

3) Stimulus: the stimulus, i.e., what the subject is required to do during the experiment, is represented by an animated hand model appearing on the PC screen situated at a comfortable distance. The model is controlled using exactly the same 6 motion values at a real-time rate of 25Hz. See Figure 2 to get an idea.

B. Data synchronisation and preprocessing

Data synchronisation is enforced on a Windows PC equipped with a multi-core processor, by gathering data from each device asynchronously and accurately timestamping each received datum. Timestamping is enforced by the HRT library [6], giving a precision of up to 1.9μs. Linear interpolation is used to find the glove motion and stimulus values best corresponding to the time at which each image is received on the PC. All data are then low-pass filtered with a Butterworth 5th-order filter, cutoff frequency at 1Hz.

C. Experimental protocol

One right-handed, male subject, 38 years old, joined the experiment. He would wear the glove and then lie his hand and part of the forearm relaxed on an orthopaedic support. Above the support, a bench vice was used to fix the ultrasound transducer just above and onto the wrist, tightly but comfortably. Standard ultrasound gel was applied between the transducer's head and the skin to allow the correct functionality of the US machine. Figure 2 shows the situation. The subject was asked to perform with his right hand what the hand model on the

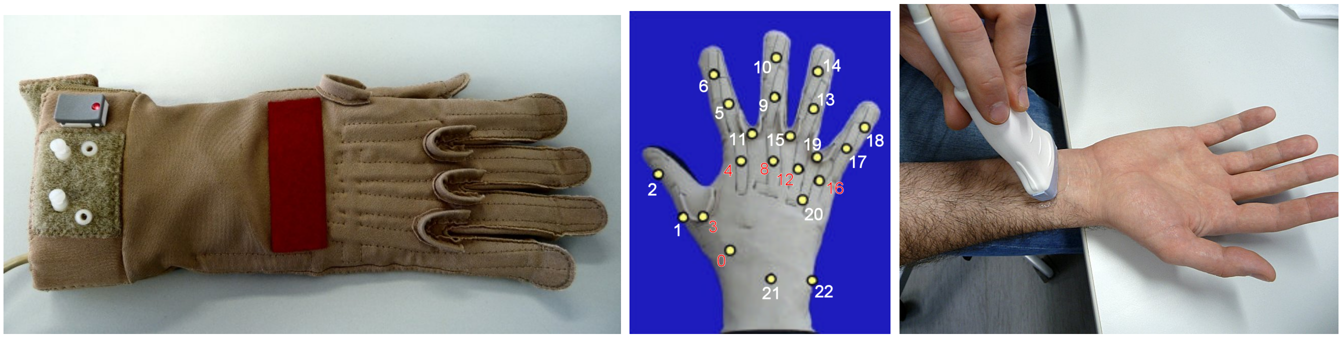

Fig. 1. Data capturing devices: (left to right) the Cyberglove; the location of its sensors (16, 12, 8, 4, 0 and 3 are used); the ultrasound transducer placed onto the subject's wrist. Moisture due to ultrasound conductive gel is clearly visible.

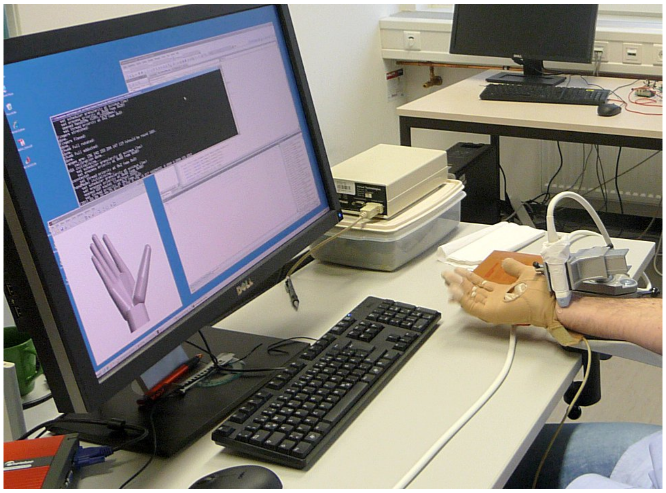

Fig. 2. The experimental setup: the subject would mimick the hand-model movements, as seen on the computer screen; meanwhile, the glove and ultrasound machine would gather hand motions and US images.

screen would perform, trying to mimick both the movement and its speed.

The stimulus consists of a sequence of basic movements, either single- or multi-finger. Single-finger movements are: pinkie, ring, middle, index and thumb full flexion and back, and thumb full adduction and back. Multi-finger movements are: (a) simultaneous flexion of the pinkie and ring, (b) simultaneous flexion of the middle and index, (c) simultaneous flexion of the pinkie, ring, middle and index, and (d) like (c) but also adducting the thumb, as in a typical "flat grasp", used to grasp credit cards or DVDs. Each basic movement is performed at three different speeds (1, 3 and 5 seconds for full flexion and back) and repeated 2 times (single-finger movements) or 3 times (multi-finger movements); inbetween movements, 1.5 seconds of rest are allowed. All in all, there are 72 movements; appropriate labels are applied to all samples in order to understand what movement and what speed is associated to each US frame and hand position. The whole experiment lasts about 6 minutes and no fatigue or discomfort was reported by the subject.

III. Image processing and feature extraction

A. *Image acquisition*

The used ultrasound machine is unfortunately not capable of delivering a stream of B-mode images directly to a PC. For this reason images have to be grabbed from standard VGA interface using a conventional framegrabber. This implies several timing problems that have to be addressed first. The ultrasound machine generates images at a rate of 28Hz. These images are sent to the VGA interface at a resolution of 1024 x 768 at 60Hz. An external framegrabber grabs these images unsynchronized at about 56 Hz and sends them to a PC under Windows via ethernet. Synchronizing this sequence of asynchronous data handling is difficult and not in the projects main focus. Therefore a proof that no frame drops occur in this processing pipeline should be sufficient. The key to this problem are three clearly differentiable noise levels in the sum of absolute differences of two consecutive images:

1) *framegrabber noise:* the same ultrasound image is grabbed twice by the framegrabber. The frame is invalid and not to be used (except in case 3, see below)
2) *ultrasound noise plus framegrabber noise:* an update of the ultrasound image occurred. The frame has changed on the US since the last grab. This valid frame is to be used.
3) *tearing:* due to the unsynchronized grabbing the top half of the image is already updated (type 2 as stated above), the bottom half not yet (type 1). The frame is invalid. The next frame of type 1 has to be used instead as a valid frame.

The image sequence passes therefore a noise dependent image classification before feature extraction marking invalid images.

Images are cropped to the valid portion of the screen showing the B-mode images and converted to gray scale.

B. *Selection of sample positions*

Features in the ultrasound image are extracted at a set of uniformly spaced sample positions M as shown in Figure 3.

C. *Feature extraction and processing*

A standard B-mode ultrasound image shows areas in the tissue where the acoustic impedance changes (such as bones or

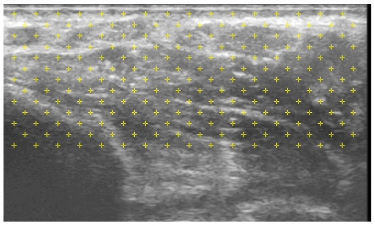

Fig. 3. Uniformly spaced sample positions with $|M| = 208$

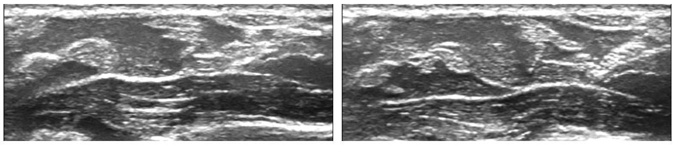

Fig. 4. Transversal B-mode sonograms of the wrist at two different finger positions

tendons) as comparably large bright regions. A superimposed noise with high spatial frequency refers to the inner structure of the tissue and is dependent on the type of tissue. Images are taken as transversal sonogram at the wrist and show a cut through the muscles and tendons at this position (see Figures 2 and 4).

Looking at image sequences from the image processing point of view while moving a finger shows some differential changes between frames and some global absolute changes:

- dominant bright structures change their shape
- dominant bright structures move
- inner tissue structure (spatially high frequency noise) shows rotational and translational movement vectors in the image plane. This inner tissue structure correlates only over a few frames since the muscles and tendons dominant movement direction is in the normal axis of the imaging plane.

1) Derivative measures: Movement direction and speed from consecutive frames can be collected easily by analyzing the optical flow. It detects the movement of tissue in the x- and y direction of the image plane. Unfortunately the dominant movement direction of muscles and tendons can not be detected. As a differential measure it will show a random bias after integration and is therefore not usable for detection of the finger positions.

2) First order measures: Interpreting the visible structures in an anatomical meaningful way is a difficult task even for highly trained doctors and can therefore not be automated by image processing. If the structures are not interpreted but modeled as interest points or edges in a reference frame, correct tracking of features in future frames is difficult and error prone due to their massively changing shape. Therefore a very simple measure is used that encodes the gray value neighborhood around each sample position. The gray value moments are calculated for each point q in a circular area with radius r around each sample point $m \in M$ with M being the set of uniformly spaced sample positions. The gray value distribution is approximated by a first order regression plane $g(r,c) = \alpha(q_r - m_r) + \beta(q_c - m_c) + \gamma$; with $g(r,c)$ being the gray value at position r,c. Therefore α denotes the mean image gradient along row direction and β along column direction respectively. Only these three features (α, β, γ) are extracted at each sample position and used for further processing. Figure 5 shows the circular region Q around a sample point m and the resulting gradient vector with the components α and β for a short image sequence at $t_0, t_3, t_6, t_9, \cdots, t_{24}$.

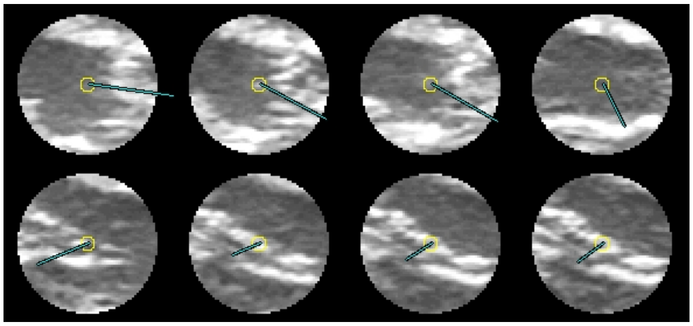

Fig. 5. Features α and β displayed as 2D-gradient vector for a short image sequence at 6.66Hz.

IV. Experimental results

The experiment detailed in Subsection II-C ended up in 7764 US frames, each one associated with a motion value obtained from the glove. At each point of the uniform grid the (α, β, γ) plane parameters are evaluated, resulting in $13 \times 16 \times 3 = 624$ image features (real numbers); the input space consists then of image feature vectors $\mathbf{v} \in \mathbb{R}^{624}$. Motion vectors (the output space) $\mathbf{p} \in \mathbb{R}^6$ consist of the 6 motion values, roughly valued[1] in $[0,1] \subseteq \mathbb{R}$.

A. Estimating K

Multivariate least-squares regression (a very basic regression technique, see, e.g., [7]) is applied to each dimension of the output space in order to obtain linear coefficients for the input space values. In other words, for each degree of motion p_j with $j = 1, \ldots, 6$, we evaluate $k_1, \ldots, k_{624}$ with $k_i \in \mathbb{R}$ such that $p_j \approx \sum_{i=1}^{10} k_i e_i$. This procedure ends up in a 4×10 matrix K, which can further on be used to estimate new motion vectors: $\mathbf{p} = K\mathbf{e}$. We employ the Matlab standard multivariate regression function.

In order to have an idea of the generality of this procedure, i.e., of how applicable this procedure is to features extracted from so-far-unseen images, we first randomly permute the data

[1] The motion range cannot possibly be strictly ensured. Calibration in the range 0-1 is performed at the beginning of the experiment by having the subject reach a few standard hand postures, e.g., full finger flexion, full finger extension etc., but nothing ensures that he won't move outside these limits now and then during the experiment.

TABLE I

ERROR RESULTS OBTAINED BY MULTIVARIATE LINEAR REGRESSION ON EACH FINGER MOTION. AVERAGE ERROR VALUES ARE DISPLAYED FOR MEAN ABSOLUTE ERROR (ERR), NORMALISED SQUARE-ROOT MEAN-SQUARE-ERROR (NRMSE) AND CORRELATION (CC).

	pinkie	ring	middle	index	th.rot.	th.add.
ERR	0.006	0.006	0.004	0.005	0.008	0.009
NRMSE	0.006	0.007	0.005	0.006	0.007	0.009
CC	1.000	1.000	1.000	1.000	0.998	0.997

set, then K is estimated on a certain subset of the data (what we will call *training set*) and tested for prediction on the remaining half (the *testing set*). Samples in the training set are normalised, as is customary, by dimension-wise subtracting the mean value and dividing by the standard deviation; with these very same statistics the testing samples are as well normalised before prediction. The prediction is repeated for 50 times (each time a different permutation), then mean and standard deviation of the obtained error are reported. As error measures, we evaluate the mean absolute error (ERR), the square-root mean-square error normalised over the range of the target values (NRMSE) and the Pearson correlation coefficient between the predicted and true target values (CC).

Table I summarises the best results, obtained when the training set consists of half of the full dataset. Standard deviations are uniformly 0 up to three digits of precision, so they are not displayed. (Recall, once again, that the ranges for the finger motions are roughly in the range $[0, 1]$.)

In order to test the resilience of this estimate to smaller training sets (or in other words, to check how many samples are necessary to obtain a reasonable estimate), we continuously decrease the size of the training set from one half to one twelfth of the data set size, and again enforce the 50-fold estimation. Figure 6 shows the results.

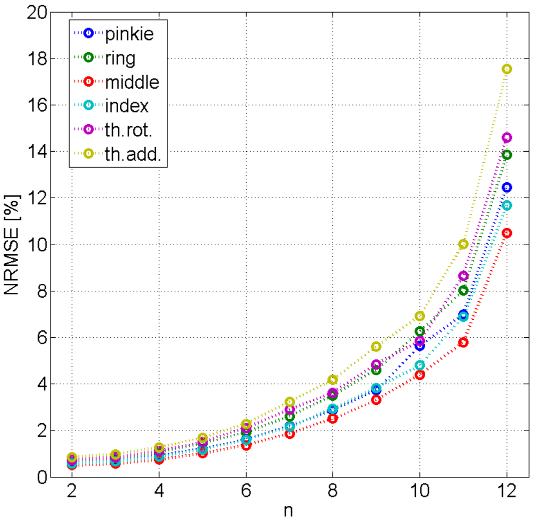

Fig. 6. NRMSE increase as the training set is progressively decreased to $\frac{1}{n}$, for all hand motions.

As one would expect, the error increases as the training set is reduced, getting as high as about 18% NRMSE in the case of thumb adduction, which uniformly remains the hardest motion to predict. As opposed to this, however, note that the error seems very resilient to decreasing training sets: for example, the NRMSE is still smaller than 5% for all motions, even if $n = 8$, that is, K is estimated over $7764/8 = 970$ samples. This means that, at least for this experiment, as few as $970/28 \approx 35$ seconds of training might be enough (recall that images are generated by the US machine at a framerate of 28Hz).

Lastly, Figure 7, upper row, shows some examples of true and predicted target values; 1294 samples are used for training in that case ($n = 6$).

B. Compositionality of hand movements

Consider now Figure 7, lower row. Here K has been estimated on single-finger movements, and multi-finger movements have been then estimated using it. (Thumb movements are not significant since they are involved in too few multi-finger movements.) The Figure shows typical pinkie, ring and middle finger motion estimations. As one can see the situation is by far worse than when using subsets of the whole data set, and nevertheless the correlation is largely preserved (the Pearson coefficient is 0.7853, 0.7342, 0.8134, 0.7361 for pinkie, ring, middle and index).

C. Local correlation

As a last test, we evaluate pairwise correlations between the pinkie movement and the feature points. For each of the 208 points, the average correlation coefficient between the pinkie movement and α, β, γ (no data filtering this time) is evaluated. Figure 8 shows the 208 coefficients so obtained, organised in 13 rows and 16 columns as is seen in Figure 3.

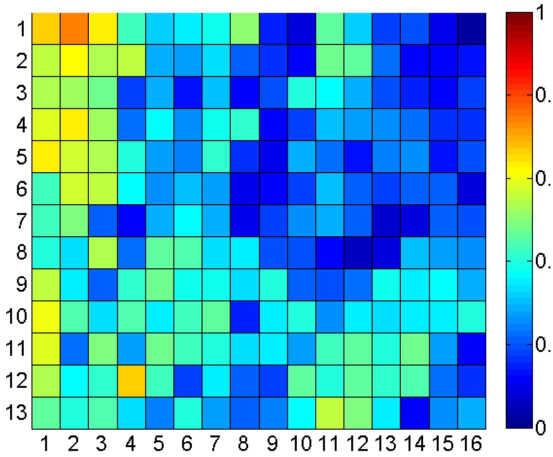

Fig. 8. Correlation between the pinkie movement and the image features. Compare with Figure 3 and example.mpg — higher correlation is apparent in the upper-left corner of the image, that is where the cross-section of the muscle which moves the pinkie is found.

As is apparent, points at the upper left corner show a higher correlation than the average, with the pinkie movement; that is where the cross-section of the *M. Flexor Digitorum Superficialis*, actuating the pinkie finger, is located in our images. Actually, the average correlation of points number $14, 15, 16, 30, 31, 32, \ldots, 206, 207, 208$ is 0.492 whereas the overall average correlation is 0.321 (Student's t-test to check that the two sets of correlations are significantly different has $p < 0.01$).

V. Conclusions and discussion

The results of the experiment hereby reported clearly show that there is a rather stable linear relationship between some ultrasound image features of the wrist and the finger movements. In particular, we set up an experiment in which a human subject would move his fingers in a principled, repeatable way; his finger movements and the ultrasound images of the cross-section of the wrist would be gathered at the same time. Later on, local features representing the image deformations at 208 uniformly spaced points would be evaluated and linearly associated to finger positions. The linear regression shows an excellent match to the true positions, even if the sample set over which it is evaluated is reduced. Moreover, a regression matrix estimated on single-finger movements only can be used to predict, with high correlation coefficients, multi-finger movements. Lastly, as one would expect, we show that high correlation exists between, e.g., the muscle associated with pinkie flexion and features extracted where in the image the cross section of that muscle is seen.

To sum up, a simple linear relationship is established between US image features and finger positions. Since the image features we have used are computationally lightweight, and that prediction using a linear model is a very fast operation, it is foreseable that this system could go on-line and work in real-time. We are actually already working on this issue, with the main application in mind, to operate the Vincent Hand (six degrees of freedom, including active thumb rotation).

Other short- and medium-term research directions include: evaluating the relative motion of the US transducer and the subject's wrist, in order to compensate the potential errors; the use of target/features correlation to understand what the most informative image zones are, finger-wise; a multi-subject analysis of applicability; a deeper investigation on single-finger motions as the sole source for the estimation of the matrix K.

The final application of this system would be, of course, that US images from the stump of an amputee could position-control a dexterous prosthetic hand or a 3D hand model on a screen, for phantom-limb pain therapeutic purposes. The applicability to amputees obviously depends on (a) the level of amputation - proximal or distal, (b) the level of residual muscle activity in the stump, (c) the level of reinnervation subsequent to the operation. Recent literature about the use of surface electromyography and TMS in such patients [8], [9], [10] lets us hope for the best.

Acknowledgements

We would like to thank Mr. Emanuel Zarka for assisting during the experiment preparation and for writing some of the data-gathering code.

References

[1] P. L. Cooperberg, I. Tsang, L. Truelove, and W. J. Knickerbocker, "Gray scale ultrasound in the evaluation of rheumatoid arthritis of the knee," *Radiology*, vol. 126, pp. 759–763, 1978.

[2] L. De Flaviis, P. Scaglione, R. Nessi, R. Ventura, and G. Calori, "Ultrasonography of the hand in rheumatoid arthritis," *Acta Radiol*, vol. 29, pp. 457–460, 1988.

[3] G. A. W. Bruyn and W. A. Schmidt, *Introductory Guide to Musculoskeletal Ultrasound for the Rheumatologist*. Bohn Stafleu & Van Loghum, 2006.

[4] J. Shi, S. Hu, Z. Liu, J. Guo, Y. Zhou, and Y. Zheng, "Recognition of finger flexion from ultrasound image with optical flow: A preliminary study," *Proc. International Conference on Biomedical Engineering and Computer Science*, 2010.

[5] B. K. P. Horn and B. G. Schunk, "Determining the optical flow," *Artificial intelligence*, vol. 17, pp. 185–203, 1981.

[6] J. Nilsson, "Implementing a continuously updating, high-resolution time provider for windows," *The MSDN Magazine*, 2004. [Online]. Available: http://msdn.microsoft.com/en-us/magazine/cc163996.aspx

[7] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data, 2nd edition*. John Wiley & Sons, Inc., 2002.

[8] C. Mercier, K. T. Reilly, C. D. Vargas, A. Aballea, and A. Sirigu, "Mapping phantom movement representations in the motor cortex of amputees," *Brain*, vol. 129, pp. 2202—2210, 2006.

[9] K. T. Reilly, C. Mercier, M. H. Schieber, and A. Sirigu, "Persistent hand motor commands in the amputees' brain," *Brain*, vol. 129, pp. 2211—2223, 2006.

[10] C. Castellini, E. Gruppioni, A. Davalli, and G. Sandini, "Fine detection of grasp force and posture by amputees via surface electromyography," *Journal of Physiology (Paris)*, vol. 103, no. 3—5, pp. 255—262, 2009.

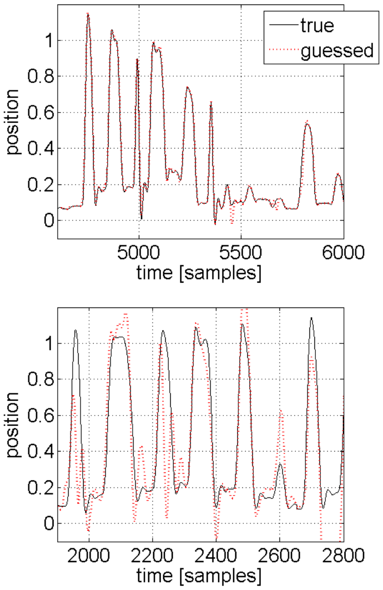

Fig. 7. (upper row) Typical true and predicted target values: (left to right) pinkie, thumb rotation, thumb adduction. The matrix K is estimated here for $n = 6$, that is using 1294 samples. (lower row) Predicting multi-finger movements using a K estimated on the single-finger movements; (left to right) pinkie, ring and middle.