UNIVERZITET U BEOGRADU
ELEKTROTEHNIČKI FAKULTET

Stereovizijski Sistem za Generisanje Percepcije

MASTER RAD

Kandidat:
Marković Marko 3119/10

Mentor:
Prof. dr Dejan Popović

Beograd, septembar, 2011.

Zahvalnica

Osnovnu ideju za ovaj projekat, kao i neophodno vođenje obezbedio je prof. dr Dejan Popović na čemu mu se zahvaljujem.

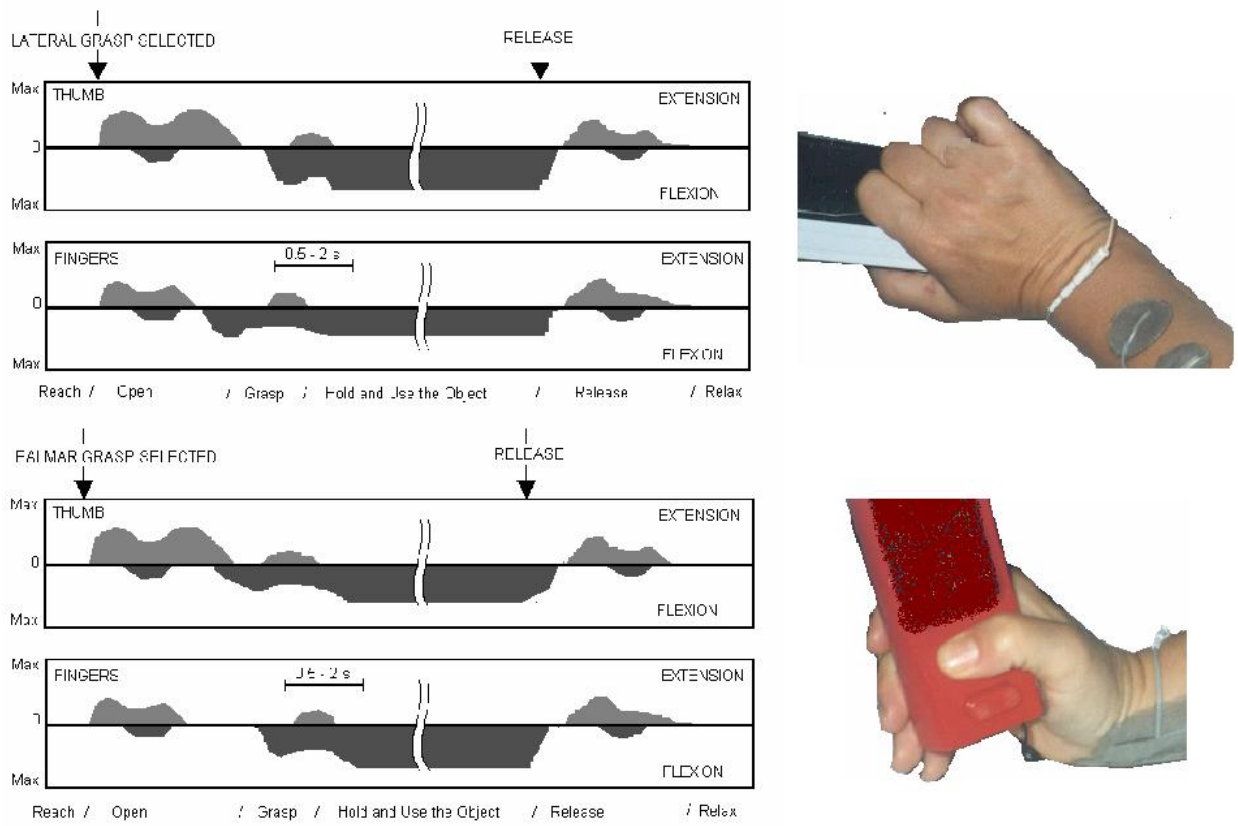
SADRŽAJ

Cilj Istraživanja

Predmet ovog istraživanja je, pre svega, pomoć osobama sa oštećenjima nervnih puteva kojima se upravlja šakom. Najčešće posledice su tremor, nemogućnost adekvatnog hvatanja predmeta, kao i psihološke posledice, koje uzrokuju povećan stepen nezadovoljstva i nervoze pacijenata, što je veoma bitan faktor za njihovo dalje normalno funkcionisanje. Dobra praksa u rehabilitaciji gornjih ekstremiteta pacijenata koji su doživeli moždani udar ili neki drugi vid neurološkog oštećenja je da se lečenju pristupi što je ranije moguće. Jedan od načina kojim se pojačava dejstvo fizikalne terapije je upotreba višekanalne električne stimulacije [1], [3], [4] na taj način omogućavajući uvežbavanje svrsishodnih pokreta (kao što su hvatanje i manipulacija različitim predmetima) koje pacijent nije sposoban voljno da izvede. Ova metoda nazvana Funkcionalna Električna Terapija predložena je još 70-tih godina 20. veka ali je na popularnosti dobila tek pojavom uređaja kao što su Bioness H200, Bionic glove i posebno Actigrip i UNAFET.

Slika 1: NESS H200 kombinuje ortozu i FES (Bioness Inc., CA).

Za efikasnu upotrebu stimulacije u okviru FET-a potrebno je prilagoditi stimulacioni program vrsti hvata koja u potpunosti zavisi od karakteristika objekta (veličine, oblika, i mase). Najčešće korišćeni (a ujedno i raspoloživi) programi stimulacije indukuju palmarne i lateralne hvatove koji se koriste u preko 90% dnevnih aktivnosti.



Slika 2: Sinergije za lateralni i palmarni hvat. Vertikalne ose su normalizovane impulsne stimulacije čija amplituda zavisi od veličine i mase objekata. Horizontalne ose su vreme.

Jedan od nedostataka dosadašnjeg načina korišćenja FET-a je što korisnici (ili medicinsko osoblje ukoliko korisnik nije u mogućnosti) moraju sami da izaberu modalitet hvata na osnovu karakteristika objekta kojim se želi manipulisati. Rešenje ovog problema predloženo je nedavno [8], [9] i zasniva se na korišćenju kompjuterske vizije radi procene dimenzija i položaja objekta, i niza *if-then* pravila na osnovu koji se potom bira prikladna strategija hvata. Unapređenje pouzdanosti sistema je usledilo ubrzo u vidu novih hardverskih rešenja [12], međutim problemi vezani za količinu informacija koje su bile raspoložive 2D pristupom u obradi slike su ostali nerešeni.

U ovom radu je predstavljen nov pristup za 3D modelovanje predmeta u kompjuterskoj viziji a sa ciljem generisanja veštačke percepcije koja je po karakteristikama bliska ljudskoj. Rad je nastao kao rezultat projekta realizovanog u okviru laboratorije za Biomedicinsku Instrumentaciju i Tehnologije (BMIT) Elektrotehničkog fakulteta u Beogradu i organizovan je u 3 celine:

- Uvod, u kome su opisane tipične oblasti primene mašinske vizije;
- Rad, integralna verzija rada koja je poslata na recenziju za časopis IEEE Transactions on Biomedical Engineering;
- Prilog, objašnjava korisnički interfejs i način upotrebe programa.

O Kompjuterskoj Viziji

Iako se često poistovećuje sa akvizicijom slike putem kamere, kompjuterska vizija predstavlja daleko širu oblast i odnosi se na rešavanje unapred definisanih zadataka obradom i izvlačenjem informacija iz slike čija akvizicija može biti izvršena sa različitih uređaja (u istoj meri kao i kamere to mogu biti i medicinski skeneri) . Ona je postala interesantan pristup za rešavanje mnogih savremenih problema u industriji, robotici, zdravstvu i zabavi u poslednje tri decenije, tj. onog trenutka kada su kompjuteri postali dovoljno moćni da se nose sa zadacima koji podrazumevaju obradu velike količine podataka.

Zadaci koji se postavljaju pred KV sisteme najčešće podrazumevaju :

- Detekciju i praćenje pokreta/objekta/oblika;
- Analizu teksture;
- Prepoznavanje oblika;
- 3D rekonstrukciju slike.

A osnovni zahtevi koje oni moraju da ispunjavaju uključuju:

- Rad u realnom vremenu;
- Robusnost u predviđenim uslovima.

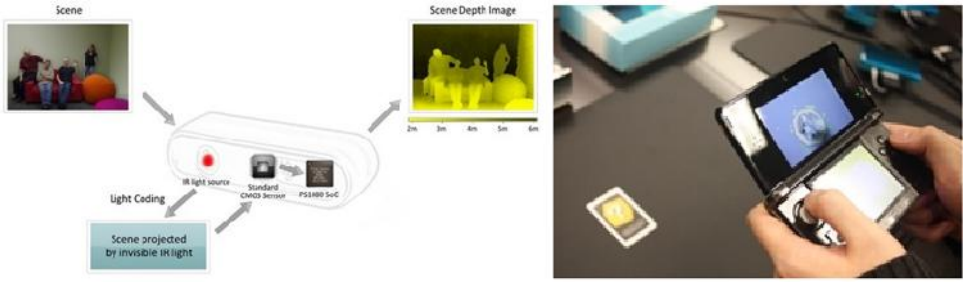
Potrebno je napraviti razliku i između sistema u kojima je potrebno donositi odluke i vršiti akviziciju u realnom vremenu i onih koji to ne zahtevaju. Kompjuterska vizija u realnom vremenu predstavlja dodatan izazov jer se pored robustnosti kao odlučujući faktor uvodi i vreme. Često se pravi kompromis između ove dve promenljive i generalno sistemi koji rade u realnom vremenu koriste brže algoritme i pouzdanije akvizicione uređaje kako bi ispunili oba zahteva.

Pored standardnih oblasti primene kompjuterske vizije, kao što su medicina i kontrola industrijskih procesa, u poslednjoj deceniji su ukazale su se i neke posebno inovativne koje će ukratko biti predstavljene.

Detekcija i praćenje objekta kombinovana sa njegovim prepoznavanjem je često korišćen koncept koji je našao primene, između ostalog, i u auto-industriji. Tako noviji modeli automobila imaju sisteme za detekciju drugih vozila, pešaka, saobraćajnih znakova, parking mesta. Algoritmi koji se ovde primenjuju podrazumevaju rad u realnom vremenu i robusnost sistema i najčešće se baziraju na detekciji specifičnih elemenata na slici (na pr. sva kola imaju farove, svi STOP znaci su crveni, ljudi imaju specifičan oblik i/ili brzinu kretanja, trake na putu su označene belom ili žutom bojom, itd.).

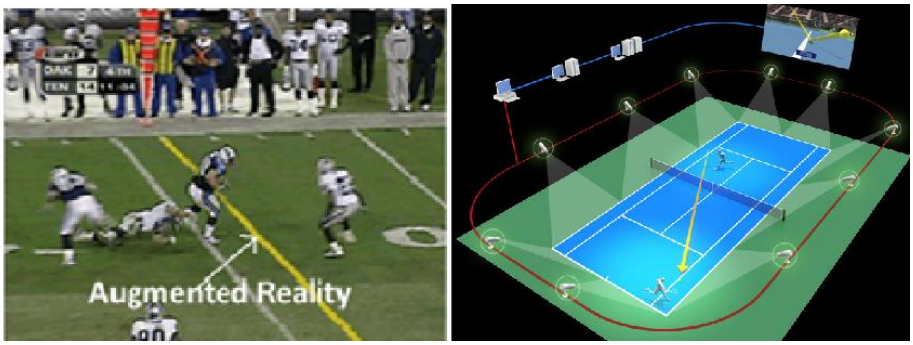
Slika 3: Volvo-ov model S60, pomoću kamere i senzora rastojanja detektuje sve učesnike u saobraćaju (levo). Detekcija napuštenog objekta u regionu od interesa, na jednoj od sigurnosnih kamera (desno).

Sigurnosne pretnje na mestima javnog okupljanja (pre svega podzemne železnice) dovele su do toga da se u sistemima sigurnosnih kamera upotrebe jednostavni algoritmi koji koriste oduzimanje pozadine za detekciju napuštenih objekata. U industriji zabave se već par godina kao igrački periferali koriste kamere koje prate pokrete i gestove igrača. U 2002. Sony je predstavio Eyetoy za svoju igračku konzolu PS2, koji je koristio jednu USB kameru niske rezolucije kako bi pratio 2D pokrete igrača. U međuvremenu je tehnologija uznapredovala pa je Microsoft 2010. napravio daleko savršeniji Kinect za Xbox 360, koji koristi IC struktuirano osvetljenje radi procene dubine, time omogućavajući 3D pokrete igrača.



Slika 2: Principijelna šema Microsoft-ovog *Kinect* sistema (levo). Nintendo je iskoristio stereoviziju i markere kako bi formirao augmentovanu realnost za svoju prenosnu konzolu 3DS i time inovirao način igranja (desno).

Izmenjena stvarnost, koja putem definisanih markera na sceni kombinuje virtuelno i realno, nekada korišćena isključivo u holivudskim filmovima, u poslednjoj deceniji nalazi primenu i u svim sportskim dešavanjima za označavanje nekih dešavanja na terenu ili za reklamiranje sponzora.



Slika 3: "First Down Line" se već par sezona koristi u NFL ligi (levo). „Hawk-eye“ (desno) sistem koristi najmanje 4 kamere da bi u realnom vremenu pratio tenisku lopticu, generisao statistiku, i omogućio igračima da dovedu u pitanje odluku linijskog sudije (jer je prosečna greška sistema oko 3 mm).

Stereovision System for Perception Generation

Marko Marković, Dejan B. Popović, *Member, IEEE*

Introduction

Functional electrical stimulation (FES) is used for the treatment of individuals whose body parts are paralyzed as a result of a neurological disorder (e.g. stroke, spinal cord injury). It has been shown that the application of FES in the treatment of upper extremities on the patients who suffered from stroke has a therapeutic effect [1], especially if the stimulation is integrated into a task oriented upper extremity exercise (e.g., exercising typical daily activities). This rehabilitation treatment requires that the stimulation paradigm is optimized for a specific grasp type (lateral, palmar, precision) and arbitrary position of the object within the workspace. Ultimately, it would be optimal if the assistive system for the rehabilitation integrates an artificial perception system that results with similar output to the one used for grasping by healthy subjects.

During the reaching to grasp process the wrist-hand complex is brought to the appropriate location in the vicinity of the object and oriented based on the type of grasp (transport or reach component), and preshaped to formation the optimal opposition space for stable grasp (prehension) [2]. Visual information provides perception about extrinsic properties of the target object. The perception about the location and orientation relative to the observer guide the transport component, while perception of the size and shape guide the preshaping of the hand [2].

This study is related to the design of an automatic procedure for the estimation of the type of grasp by the stereovision system (SVS). It is anticipated that once the type of grasp and position of the object are known the appropriate sequence of stimulation of the upper arm, forearm, and hand muscles would bring the hand into the appropriate position and that the hand opening and closing will be part of the near normal grasping [3]. The camera system is planned to be used as an interface for humans and the FES grasping assistance to select the method appropriate electrical stimulation paradigm during the therapeutic session, and possibly in orthotic applications [4].

The basis for this approach follows the suggestions that there are common denominators in the shape, size, and position of the object to be grasped used by most humans [2]. In other words humans have tendency of generalizing access strategy for similarly shaped objects. We tested this idea and performed a study where we showed that different subjects have similar approach in prehension of same objects (Table I) [5].

In robotics general approach for object manipulation and pose estimation, is to create object-specific model (planar - 2D or spatial - 3D, depending on a purpose) and use it as a reference. In recent years many algorithms [6], [7] have successfully addressed this problem, but none of them had the ability of mentioned human-alike generalization. The task of our research is somewhat different since the application of FES allows the user to volitionally make

adjustments by changing the body posture; thereby, eliminate the shortcomings of the simplified solution that we suggest.

The proposed artificial perception has been originally tested with transradial prostheses [8], [9]. In these studies a single camera system has been used to estimate the 2D position and type of grasp. A more detailed presentation of the methods used for the estimation of the position and size of objects can be found in [10], [11]. Since the object analysis was only for the 2D space, the system was unable to discriminate between different poses of the same object. Furthermore 2D analysis has major disadvantages when compared with the 3D analysis in the richness of information. We present a novel method which combines 2D and 3D image processing.

This approach has been tested within the FES technology, and pilot results demonstrated that it can operate for simple objects positioned at the specific location in the real time when using the PC platform and MATLAB program [12].

Material and Methods

HARDWARE IMPLEMENTATION

Every image acquisition system, either the human or machine visual system, by its nature performs some kind of transformation of real 3D space into 2D local space. In field of computer vision number of techniques (e.g. passive or active stereovision) and technical solutions (e.g. TOF or structured light cameras) have been developed to address this problem (i.e., getting depth information). Since it is technically most available we have adopted passive stereovision approach, by using custom made machine vision system embedded in a small, lightweight plastic box. The complete hardware system (Fig. 1) consists of:

1. *Two identical USB CMOS web cameras, operating at resolution of 640 by 480 pixels, aligned in a manner that approximately meets canonical stereo configuration setup [13]. The stereo baseline is 6.2 cm and focus on both cameras is set to reproduce sharp images at distances between 30 cm and 150 cm;*
2. *One red laser diode placed in between two cameras so that projected laser beam is parallel with the optical axis of camera lens. The pointing diode shares the power supply with one of the cameras and is operated programmably in a switched on and off manner via RS-232 connection. Essential purpose of the pointing laser diode is to mark object of interest thus providing visual feedback to the user on the cameras orientation and successful object detection (laser diode is switched off by the algorithm);*
3. *One inertial sensor (3-axis, ± 2G range accelerometer) placed on a hand, designed for measuring tilt of the wrist-elbow system (since amount of hand rotation is relative to its initial orientation).*

SVS is designed to be integrated into a light hat or some similar head mounted device so that the cameras and laser are oriented in the direction of the working space in which the objects are located (approximately 0.8 radians down to the horizontal plane), as shown on Fig. 1. This approach is different to the *eye-in-hand* method described in [8], which was developed for an intelligent artificial hand. The proposed concept can be viewed as quasi-stationary, because the cameras move only when the subject decides to use another object, and triggers the operation.

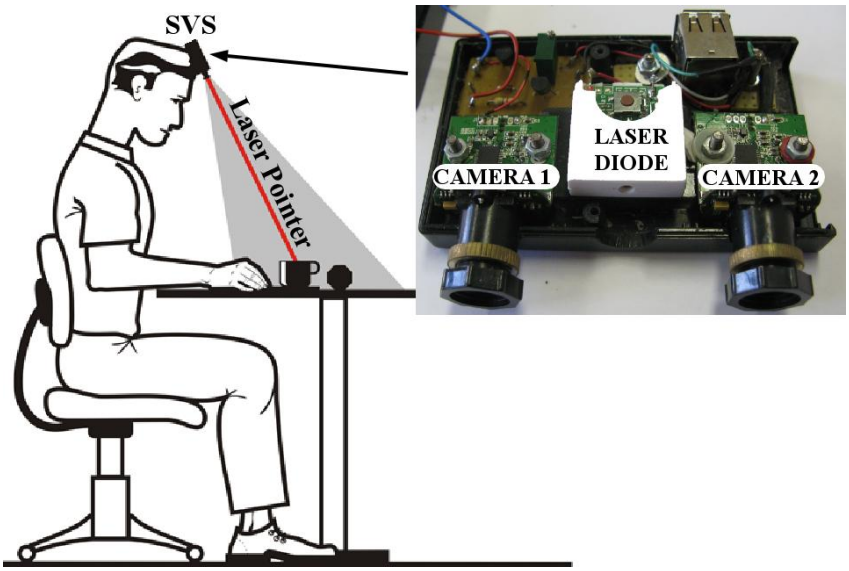


Fig. 1. Proposed system setup (left) and its hardware implementation (right). During acquisition phase system operates in real time at 15 fps, whereas laser diode pointer discriminates object of interest from others that are in SVS FOV (Field Of View).

SYSTEM CALIBRATION

Rectification transformation parameters of stereo images as well as intrinsic and extrinsic properties of previously rectified stereo camera configuration were found via MATLAB Camera Calibration Toolbox [14] using set of 20 chessboard pattern images (obtained from different relative camera/pattern positions). By combining intrinsic camera properties with weak perspective projection approximation [15] 3D image can be reconstructed if relation between pixel disparity and depth (i.e., distance of the point to the camera) is previously found using (1).

$$z_{world} = focal_{length} \cdot \left(1 + stereo_{baseline}\right) / disparity \quad (1)$$

$$z_{world} = a + b / disparity \quad (2)$$

If the distance between two cameras is not well known (in our case it cannot be measured with absolute accuracy) a general form (2) of (1) is preferred for calculating depth from disparity. We can solve for the two unknowns (a, b) via least squares by collecting a few corresponding depth and disparity values from the scene and using them as tie points.

Whereas SVS calibration is one time job, the accelerometer needs to be recalibrated (where calibration process in this case refers to using sensors initial output as referent) upon every new session.

GRASP PARAMETERS ESTIMATION

In clinical application, transradial FES induces hand movements that fall in one of the three distinct grasp strategies: lateral, palmar, and pinch, thus this research is focused on evaluating two main parameters: grasp type with appropriate wrist-elbow rotation, and grasp size

First parameter was determined heuristically by analyzing different grasp strategies on 10 healthy right-handed subjects (5 male and 5 female, average 23 year old) during simple object manipulation task (Table I). In a similar fashion, by measuring finger flexions (using three goniometers placed on thumb, index and middle finger) in various scenarios during grasping phase (i.e., different object dimensions) we can use linear or cubic regression model to describe dependency between object dimensions and resulting grasp size [5].

TABLE I
DISTRIBUTION OF EMPLOYED GRASP STRATEGIES FOR DIFFERENT OBJECTS

Item	Item description and placement	Palmar grasp	Lateral grasp	Pinch grasp
Cup	Average sized, placed so that handle is not easily reachable	90 %	10 %	0 %
CD box	Standard sized (non-slim), placed vertically	0 %	80 %	20 %
Bottle	Average sized – 0.5 l, placed vertically	100 %	0 %	0 %
Mobile phone	Average sized, lying on the table surface	0 %	0 %	100 %
Spoon	Metal soup-spoon, lying on the table	0 %	90 %	10 %
Pencil	Average sized, lying on the table surface	0 %	80 %	20 %
Juice box	Large – 1 l, placed vertically	100 %	0 %	0 %

For every object separately the most frequent grasp type is declared as relevant one and serves as a reference approach for all other objects that are similar in shape with the original (e.g., for all bottles, no matter what their size is we can assume palmar grasp strategy).

OBJECTS DATABASE

As object analysis cannot be performed without any a priori knowledge, a simple yet effective modeling technique is proposed. Instead of trying to determine specific approach for every object separately, we divide them into groups that have same global features (e.g., all cups are described with conic model and employ palmar grasp strategy; all boxes are described by planar surfaces and also employ palmary grasp strategy, etc.). Every model in a database, is thus described by: object primitive (i.e., binary, 90 pix x 90 pix, image that describes basic object shape), a pair of grasp related points (reference points on object primitive), and a type of regression method (used for 3D object modeling).

ALGORITHM

Algorithm consists of three distinct phases, which execute in an infinite loop in alternating fashion as shown in Fig. 2. In situations where there is low certainty that made decisions are correct it is better to skip FES stimuli than forward parameters for the inappropriate one. With that in mind program flow is periodically interrupted to check if output data conforms to predefined constraints (e.g., surface on which objects are placed must be planar, estimated grasp parameters must conform to physical constraints, …).

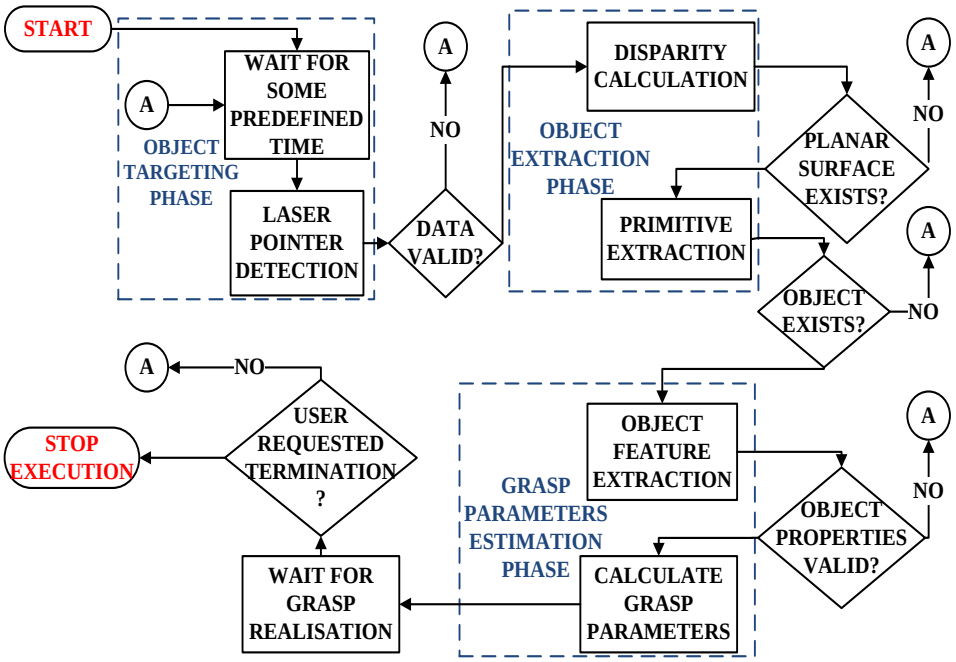


Fig. 2. Algorithm execution flow chart. Program works in an infinite loop until user requests termination, whereas any inconsistency in output data results in current execution to be reset to the object targeting phase.

Laser Detection

Laser detection is done only in a specific ROI (Region of Interest) since laser diode is placed in between cameras (as it is described in hardware implementation section). The algorithm first performs a color space conversion from original RGB image - IM_{RGB} to YCbCr color space image, and then analyses all pixels in [Y] and [Cr, Cb] components separately to form binary image BW_L as described further below.

$$BW_Y(i,j) = \begin{cases} 1, & Y_{ROI}(i,j) \geq T1 \\ 0, & otherwise \end{cases}$$

$$BW_{Cr}(i,j) = \begin{cases} 1, & 1.5 \cdot Cr_{ROI}(i,j) - 0.45 \cdot Cb_{ROI}(i,j) \geq T2 \\ 0, & otherwise \end{cases}$$

$$BW_L(i,j) = BW_Y(i,j) \wedge BW_{Cr}(i,j)$$

$$T_1 = 0.85 \cdot \max([Y_{ROI}])$$

$$T_2 = 0.85 \cdot \max([1.5 \cdot Cr_{ROI} - 0.45 \cdot Cb_{ROI}])$$

$$i = 1,2,\ldots,\dim(Y_{ROI},1),\ j = 1,2,\ldots,\dim(Y_{ROI},2)$$

*$\max([X])$ is maximum value in matrix X

Due to errors and/or existence of specular reflections (especially on red surfaces) there can be more than one 8-connected group of pixels in BW_L. This problem is addressed by using specific morphological features of a laser mark (i.e., it should be similar to a circle) to decide best circle-fit group, and then the final laser tag pixel location (lt_x, lt_y) is calculated as a median value of all pixel locations that belong to the best-fit group.

Disparity Map Calculation

Depth estimation (i.e., disparity map calculation) from binocular imagery represents a crucial step, which success rate has major impact on quality of all further image processing phases. Efficient Large-Scale Stereo Matching algorithm as described in [16] is employed on a grayscale image pairs to deal with this task. As well as with all stereo matching algorithms two main problems arise while solving pixel correspondence problem:

1. *"Similar" instead of "Same" - leading to local errors and discontinuities in disparity estimation;*
2. *Some parts of the scene are visible only in one camera (i.e., occlusion problem) - leading to unresolved disparity values;*

Further step involves 3D image reconstruction by using parameters and procedures described in calibration section.

Modeling Table Surface

Planar model of a surface (on which objects are placed) is assumed on the basis of specific scene setup. RANSAC [17] method is used to assessment surface equation $(ax+by+cz+d=0)$ parameters by finding its three support points in 3D image matrix - IM_{3D} that have largest number of inliers (where inlier is every point that has distance to the plane less than 0.3 cm). This robust and outlier-suppress effective technique is implemented by randomly choosing N (e.g., 1000 or more) non plane-degenerate (i.e., not mutually too close and not collinear) triplet points that are uniformly distributed along IM_{3D}, and then counting number of inliers for every plane defined by these triplets. Finally, logical matrix BW_T containing all inliers locations of best-fit table surface plane is formed.

Object Primitive Extraction

Under the assumption that table has approximately unicolored surface, object primitive BW_O is extracted by means of Mahalanobis distance-based color segmentation [18] and gradient based post processing. Initial object primitive is found by combining laser tag location (lt_x, lt_y) with either BW_T or BW_{Tc} (initial color segmented table surface – used if object detection in BW_T fails). Pixels required for calculating Mahalanobis distance parameters in any given image region (IM_{ROI}) are given by:

$$P_S = \left\{ pix \in IM_{ROI} : \left| pix(i,j,k) - C_{med}(k) \right| < T \wedge \sum_{k=1}^{3} \left| pix(i,j,k) - C_{med}(k) \right| < 2T \right\}$$

$C_{med}(k)$ is median value of every color component in IM_{ROI}

$$T = (0.1 \div 0.2) \cdot \max\_color\_value$$

$$i = 1,2,\ldots,\dim\left(IM_{RGB},1\right),\ j = 1,2,\ldots,\dim\left(IM_{RGB},2\right) \qquad (3)$$

$k \in \{1,2,3\}$ – equivalent to R, G, B color components

After initial segmentation, due to need of robust and precise primitive extraction, two-step post processing is applied on object primitive and its corresponding RGB image:

1. *Color ROI-based post processing – divides bounding box of object primitive into overlapping regions (as shown in Fig. 4.3). Color segmentation is then applied in a specific order, with segmentation thresholds that become more restrict as regions approach center of original primitive;*
2. *Gauss gradient based post processing – finds and process gradients of primitive edge pixels* G_E *, along all color components separately. Gradients are compared against two thresholds, thus forming two groups of pixels:*

$$G_{weak}(k) = \{pix \in G_E(k) : pix < T_l(k)\}$$

$$G_{med}(k) = \{pix \in G_E(k) : pix \geq T_l(k) \wedge pix < T_h(k)\}$$

$$\mathrm{T}_l(k) = \tilde{G}_T(k), \mathrm{T}_h = \tilde{G}_E(k)$$

G_T -gradient values of table surface inliers

$k \in \{1,2,3\}$ - equivalent to R, G, B color components

Every 8-connected $G_{weak} - G_{med}$ *(in all three color components) group that has enough weak support points is than removed from object primitive and new edge layer -* G_E *is found. The process is repeated, until number of found weak-edge pixels falls below certain threshold.*

Primitive Matching

After extraction process, by using (4) as a measure of similarity, object primitive is compared against all reference primitives in a database, with aim to find a pair with highest similarity value. In this process primitive is cropped (to dimensions of its bounding box), resized (to resolution of 90 pixels along longest axis), translated (via FFT based template matching) and rotated (with variable step size) until best match is found (Fig. 4.5).

$$S(k) = \left(\sum_{i=1}^{M} \sum_{j=1}^{N} BW_O^*(i,j) \cdot BW_{B_k}^*(i,j) - \sum_{i=1}^{M} \sum_{j=1}^{N} \left| BW_O^*(i,j) - BW_{B_k}^*(i,j) \right| \right) / \sum_{i=1}^{M} \sum_{j=1}^{N} BW_{B_k}^*(i,j) \quad (4)$$

* Image pixel is converted from logical to integer data type

$M = \dim(BW_O, 1), N = \dim(BW_O, 2), (i, j)$ are defined in (3)

BW_{B_k} – k-th reference primitive from database

Finding Grasp Related Object Patch

Grasp relevant points from best-match reference image are mapped onto transformed (i.e., cropped, resized, and rotated) object primitive, by maintaining constraints that define them (i.e., direction of line passing through points remains unchanged, and their relations to the object primitive are the same as those to the reference primitive). Inverse transformation process (i.e., rotation, resizing, and padding) of this newly formed image results in locations of grasp relevant points that are in reference frame of original object primitive.

Grasp related object patch BW_{Op} is defined as every subset of object pixels (Fig. 4.6) that forms a line (L_k) parallel with the reference one (i.e., line that is defined by two previously found reference points) and conforms to these requirements:

1. *Length of line (in pixels) needs to be similar to the original one;*

2. *Number of line endpoints that lie on object boundary is greater or equal than number of the boundary endpoints on the original line.*

Matrix of 3D points P_{3D} that describes object patch in 3D space is given by (5).

$$P_{3D}(l,k,n) = \{ p(i,j,n) \in IM_{3D} : BW_{Op}(i,j) = true \wedge BW_{Op}(i,j) \in L_k \} \quad (5)$$

$N = \min([\dim(L_1),...,\dim(L_M)])$, M - number of found 2D lines

(i, j) are defined in (3) and $n \in \{1,2,3\}$

$k = 1,2,...,M$, $l = 1,2,...,N$

Fitting Model to Object Patch in 3D Space

Properties of object patch (i.e., its pose estimation relative to the table surface and SVS) needed for grasp estimation, are estimated by means of regression. Every object type is described using different mathematical model, which basic properties are a priori defined in a database.

Due to gross errors that appear in disparity matrix and limited precision of object primitive extraction, P_{3D} will have moderate number of points that can be considered as noise. It is assumed that object patch can be described by N close lines that are similar in orientation. These lines were found by employing RANSAC 3D line fitting [19] column-wise on P_{3D} , thus creating new filtered set P_{3DS} (Fig. 4.6) which only contains those points from P_{3D} that form lines with compatible orientations (where reference orientation is defined as median value of all lines orientations and is also used as plane normal for reference intersection plane) and have enough supporting points (inliers).

Specifically, depending on object size and shape (i.e., matched primitive from base), one of three regression models is used:

1. *Box model –Iteratively fits planar surfaces using RANSAC algorithm [19] until it finds two planes with compatible orientation (i.e., they must intersect and form an angle that is close to 90º). Using these two planes intersection line direction and subset of inliers in P_{3D} cloud, 3D edges are reconstructed and semi-closed 3D patch structure is formed (Fig. 3.1b and Fig. 3.2b). Depending on object relative position to the SVS and quality of disparity map, special case when only one object plane is visible is also treated as valid. In this case, since direction of edges is unknown, it can only be assumed either by intersecting plane with table surface (only valid if planes form an angle that is close to normal) or by using reference intersection plane direction found in previous step. Direction of axis, in general case, is determined with two center points of undisclosed patches, whereas its origin lies in between them. In special (degenerate) case main axis has same direction as a reference orientation, whereas its origin lies in the center of the patch.*
2. *Conic model –Every midpoint of M lines in P_{3D} is combined with intersection plane normal thus creating a set of M planes. For every plane a set of accompanying points is defined as those points in set P_{3DS} whose distance from the plane is below certain threshold (0.15 mm). These points are then projected on every intersection plane separately, and Pratt's method [20] is used to fit M circles trough them (since we have assumed conic model these points will usually form an arc with some percentage of noise*

added by irrelevant object details). By describing conic models with all possible non-degenerate (i.e., all circle pairs that have a very different radius in close distances are degenerate) and non-repeated combinations of two fitted circles and counting number of their inliers in P_{3Ds}, model with best data-fit is chosen (Fig. 3.3b and Fig. 3.4b) . Two center points of the model bases determine main axis direction, whereas point between them defines axis origin.

3. *Line model- When object patch cannot be modeled by two previously described methods, either because its form is irregular or it is too small (e.g., thin pen has radius that is close to average SVS error) a simple line model is used. This method searches all lines that were used to form P_{3DS} to find the two most distant lines that are also nearly parallel. These two lines with its inliers endpoints are used to describe a simple planar surface model, which axis is defined as in special case of a Box model.*

Finally, object pose is evaluated as orientation of its main axis in regard to the table surface model. Similarly, object dimension of interest is found in regard to camera reference system, by analyzing intersection properties between object model and a plane that has origin and orientation of a main axis.

Results

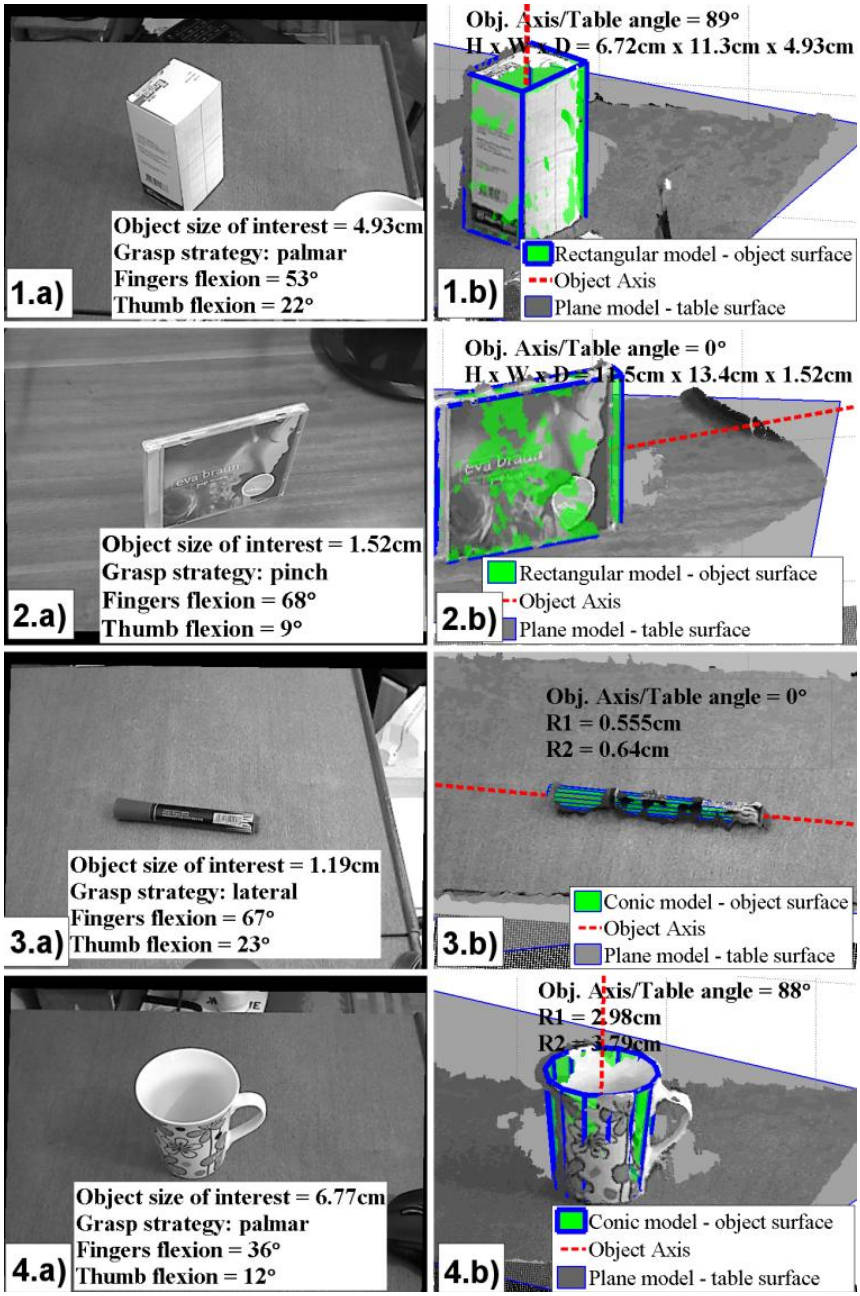


Fig. 3. Estimated grasp parameters (a) and regression models (b). Three regression models are considered in final phase: cone, plane, and series of collinear lines. Modeling is applied only on object surface patch that is of interest in regard to its grasping and manipulation (i.e., only part of the object surface is described by a model).

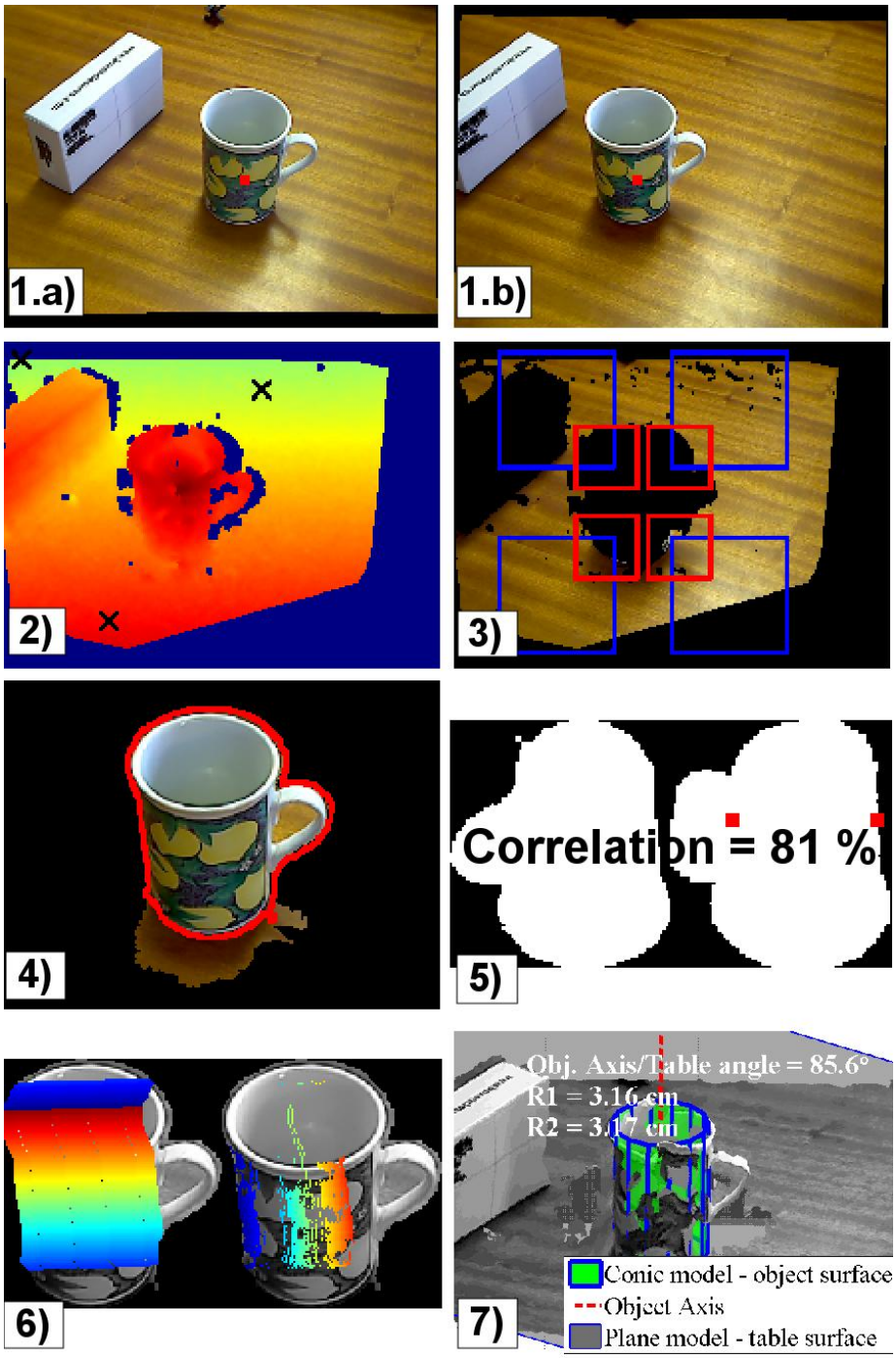


Fig. 4. Seven major steps in algorithm execution. 1) Rectified stereo image pair with corresponding laser tag (marked with red squares) locations; 2) Calculated disparity map with three support points (marked with X) that define table plane; 3) Only pixels that belong to flat surface are used for object extraction process. Inner(red) and outer(blue) ROI are used for color image segmentation (between every adjacent pair of ROI three uniformly distributed and overlapping sub ROI's of the same size are also used, but are not shown for sake of image clarity); 4) Extracted object before and after gradient based post-processing; 5) Best match between extracted and database primitive (red dots signify previously defined reference points); 6) Primitive based object patch (left image-half, every color represents a different 2D line) and inliers that form lines with compatible 3D orientation (right image-half shown in 2D space); 7) Fitted object patch model in 3D space.

The method proposed in this paper has been implemented in MATLAB environment on a commercial laptop computer (Intel Core i5 @2.7 GHz, 4 GB RAM) running under Windows 7 64-bit operating system.

Simulation of real working conditions was performed by mounting SVS on a camera tripod (thus simulating proposed head-mounting configuration) and using randomly situated accelerometer (thus simulating initial position of wrist-elbow system). Scene setup during the testing phase was typical for the proposed application and included existence of reflections and shadows of moderate intensity in a well lit room. Test was performed on 20 different (in size, shape and application) commonly used objects that were distributed along flat table surface in a manner that there is no overlapping between them in perspective of a camera system.

Representative end-results in different scenarios are given in Fig. 3, whereas illustrations of output results from previously described algorithm sections are given in one typical scenario (Fig. 4.). Average time needed for estimating object grasp parameters, (i.e. fitting object model to cloud of 3D points) from the moment when the laser pointer is successfully identified, is between five and fifteen seconds, with calculated object dimensions margin of error around 0.4 cm (commonly between 0.2 cm and 0.3 cm). It should be noted that both average computational time and margin of error largely depend on object size, complexity, and quality of previously calculated disparity map.

Discussion

Compared to the *eye-in-hand* used in [8], [9] suggested system placement has two main advantages:

1. *Analysis of principally stationary images (as opposed to motion-blurred images caused by hand tremor that is to some extent present in most patients).*
2. *More intuitive and easier usage due to head-mounting approach.*

Image segmentation performs well in different scenarios and is generally robust against existence of reflections and shadows of moderate intensity. Weak point of presented approach is that every connected pixel group that is different from background is considered to be a single object (i.e., system is unable to discriminate overlapping objects in image). Object primitive matching phase can tolerate small segmentation errors and is accurate as long as database primitives have moderate differences in shape. If previous steps were done correctly, grasp related object modeling is able to robustly fit desired model in a presence of many local inconsistencies and gross errors (caused mostly by bad disparity information).

Although less powerful than those SIFT-based [6], our method has the advantage of high level generalization (whole group of similar shaped objects can be described with a single, or very few models) that is one of the prerequisites for any application of this type. Namely, the SIFT based estimation of the pose would require the training for every new object introduced into the therapy, which is not the case with our algorithm. Ultimately the SIFT based estimation would lead to very large and redundant database. This is opposing our goal to perform one time training for every object group with similar shape and application, thus enable simple system application.

Resolution of evaluated grasp parameters fits very well with the precision that can be controlled with current FES assistive technologies; thus, generated errors can be ignored as long as they do not fall in area of gross errors (e.g., errors that can cause heavily modified grasp approach).

The testing was done in the laboratory conditions with the PC platform and running the MATLAB program. We are in the process of finalizing the C++ which will be running on the microcomputer that is integrated into the portable camera system (Fig. 1).

The proposed camera based could be integrated with the inertial sensors at the arm and hand in order to apply it for control of the elbow joint, pronation/supination and wrist rotations. The inertial system needs to output the relative position of the forearm and hand with respect the camera position (orientation).

References

[1] D. B. Popović, M. B. Popović, T. Sinkjær, A. Stefanović, L. Schwirtlich "Therapy of Paretic Arm in Hemiplegic Subjects Augmented with a Neural Prosthesis: A Cross-over study," *Canadian Journal of Physiology and Pharmacology*, vol. 82, pp. 749-756, August 2004.

[2] C. L. McKenzie, T. Iberall, *The Grasping Hand*. 1st ed., North Holland: Elsevier Science, 1994, ch. 2 to 5.

[3] D. B. Popović, T. Sinkjær, *Control of Movement for the Physically Disabled*. 1st ed., London: Springer, 2000, pp. (ili ch.)

[4] D. B. Popović, T. Sinkjær, M. B. Popović, "Electrical stimulation as a means for achieving recovery of function in stroke patients," *Journal of NeuroRehabilitation*, vol. 25(1), pp. 45-58, August 2009.

[5] N. Babić, M. Marković, "Analysis of fingers flexion during simple grasping tasks," report on school project, Scholl of Electrical Engineering, University of Belgrade. Available:

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91-110, February 2004.

[7] A. Collet, M. Martinez, S. S. Srinivasa, "The MOPED framework: Object recognition and pose estimation for manipulation," *The International Journal of Robotics Research,* vol. 30(10), pp. 1284-1306, September 2011.

[8] S. Došen, D. B. Popović, "Transradial Prosthesis: Artificial Vision for Control of Prehension," Artificial Organs, vol. 35(1), pp. 37-48, Jul 2011.

[9] S. Došen, C. Cipriani, M. Kostić, M. C. Carrozza, D.B. Popović, "Cognitive vision system for the control of a dexterous prosthetic hand: An evaluation study," *Journal of NeuroEngineering and Rehabilitation*, pp. 7-42, August 2010.

[10] M. B. Popović, D. B. Popović, T. Sinkjær, A. Stefanović, L. Schwirtlich, "Clinical Evaluation of Functional Electrical Therapy in Acute Hemiplegic Subjects," *Journal of Rehabilitation Research and Development*, vol. 40(5), pp. 443-454, September 2003.

[11] Đ. Klisić, M. Kostić, S. Došen, D. B. Popović, "Control of Prehension for the Transradial Prosthesis: Natural-like Image Recognition System," *Journal of Automatic Control*, University of Belgrade, vol. 19, pp. 27-31, August 2010.

[12] S. Došen, G. K. Kristensen, B. Bakhshaie, M. Pizzolato, M. Smondrk, J. Krohova, M. B. Popović, "Computer vision for selection of electrical stimulation synergy to assist prehension and grasp*," in Proc. 15th annual IFESS conference* ,Vienna 2010.

[13] B. Cyganek, J. P. Siebert, *An Introduction to 3D Computer Vision Techniques and Algorithms*. 1st ed., West Sussex: John Wiley & Sons, 2009, pp. 36-38.

[14] J. Y. Bouguet, "Camera Calibration Toolbox for Matlab," Computational Vision at the California Institute of Technology. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/

[15] B. Cyganek, J. P. Siebert, *An Introduction to 3D Computer Vision Techniques and Algorithms*. 1st ed., West Sussex: John Wiley & Sons, 2009, pp. 27-28.

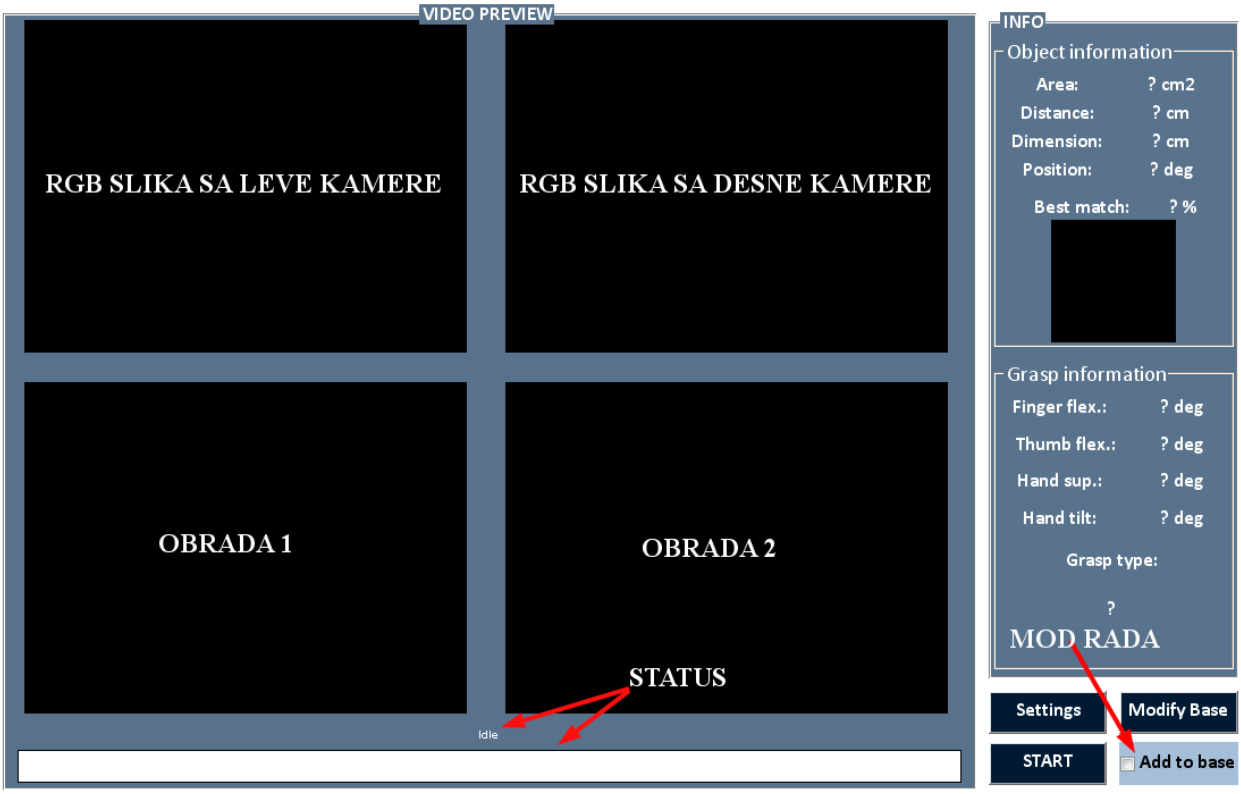
[16] A. Geiger, M. Roser, R. Urtasun, "Efficient Large-Scale Stereo Matching," in *Proc. 9th annual Asian Conference on Computer Vision*, 2010.
[17] M. A. Fischler, R.C. Bolles "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24(6), pp. 381-395, June 1981.
[18] R. C. Gonzalez, R. E. Woods, S. L. Eddins, *Morphological Image Processing in Digital Image Processing using Matlab*. 1st ed., Pearson Prentice Hall, 2004, pp. 237-240.
[19] P. Kovesi, "MATLAB and Octave Functions for Computer Vision and Image Processing," School of Computer Science and Software Engineering, University of Western Australia. Available: http://www.csse.uwa.edu.au/~pk/Research/MatlabFns/index.html
[20] V. Pratt, "Direct least-squares fitting of algebraic surfaces," *ACM SIGGRAPH Computer Graphics*, vol. 21(4), pp. 145-152, July 1987.

Prilog

KORISNIČKI INTERFEJS

Radi jednostavnosti upotrebe i testiranja razvijen je intuitivni korisnički interfejs koji se sastoji iz 3 celine (Glavni panel, Podešavanja, i Pristup bazi).

Glavni panel

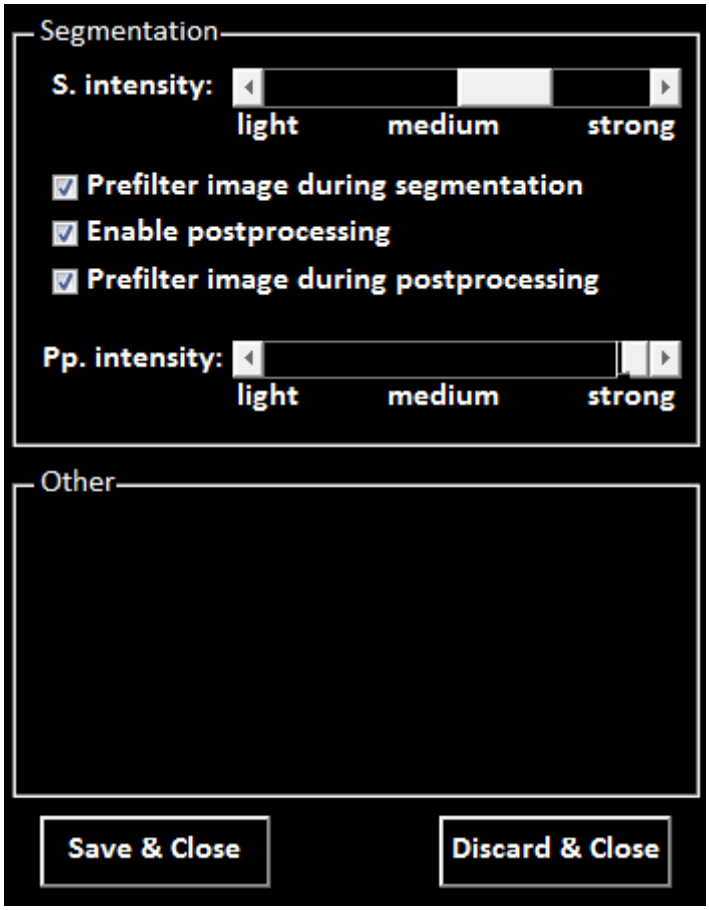


Obrada podataka se pokreće i zaustavlja pomoću jedne kontrole (START/STOP) i za vreme njenog trajanja korisniku je onemogućen pristup bazi. Program poseduje dva načina funkcionisanja:

- Standardni, koji podrazumeva računanje svih neophodnih parametara za procenu hvata;
- Akvizicioni, koji podrazumeva da se obrada podataka vrši samo do onog dela koji je potreban da se objekat (odnosno njegov primitiv) doda u bazu. Ovaj mod dakle služi kako za formiranje baze tako i za njeno opciono proširivanje.

U glavnom panelu je omogućeno praćenje postupka akvizicije i obrade slike u realnom vremenu kao i pristup ostalim celinama koje su bitne za funkcionisanje sistema. Podpaneli *obrada1* i *obrada2* zajedno sa statusnom linijom i *info* panelom služe za prikaz svih bitnijih rezultata koji proističu iz obrade podataka i time pružaju feedback korisniku o ispravnosti, i trenutnoj fazi rada kao i eventualnim greškama (svaka greška ima jedinstveni tekstualni identifikator koji se ispisuje iznad statusne linije ukoliko situacija to zahteva) . Sve informacije o objektu i parametrima hvata popunjavaju se postepeno kako obrada odmiče, a po započinjanju svakog novog ciklusa vrši se njihovo automatsko brisanje.

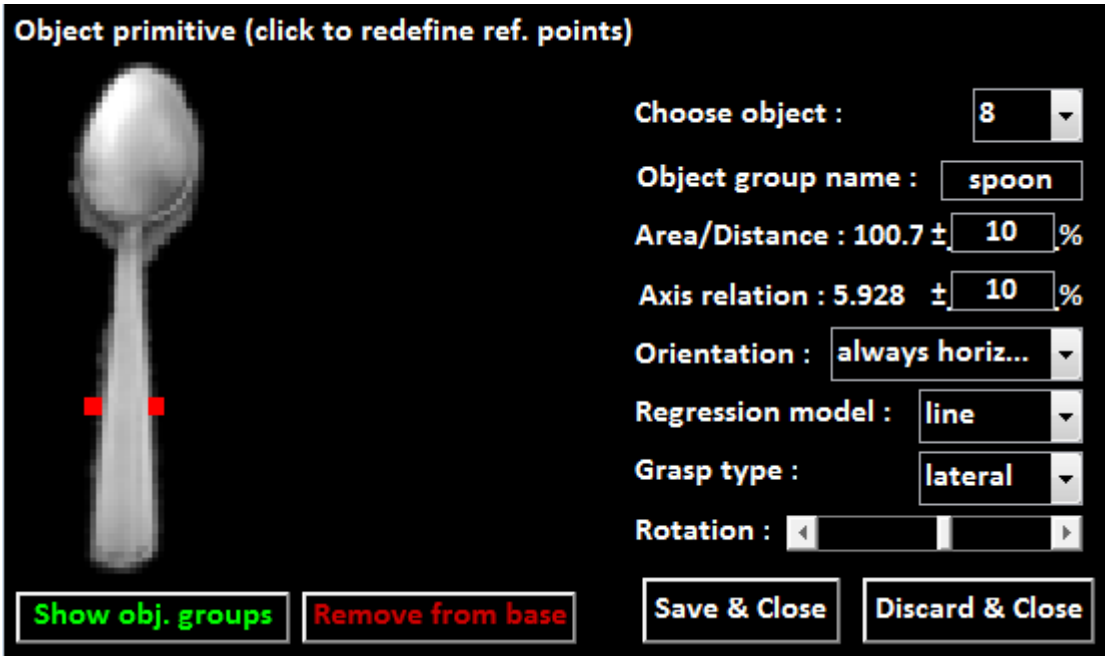
Podešavanja



Podešavanjima se pristupa iz glavnog panela u bilo kom trenutku (bez obzira da li je obrada u toku ili ne). S obzirom na to da je algoritam za segmentaciju slike osetljiv na promenu uslova rada (osvetljenost scene, refleksije, senke) u podešavanjima se nalaze kontrole pomoću kojih se vrši manipulacija njegovim parametrima, pa je korisnik u mogućnosti da iz par pokušaja pronađe kombinaciju postavki koja najviše odgovara njegovom konkretnom problemu. Pritom treba napomenuti da filtriranje prilikom segmentacije ima zadatak da omekša ivice i smanji šum kako bi segmentacija po boji bila što uspešnija, dok filtriranje pri post-procesiranju ima suprotan efekat (pojačava ivice a donekle i šum).

Sva podešavanja se upisuju na disk i učitavaju se sa pokretanjem programa (u slučaju nepostojanja fajla učitavaju se uobičajene postavke). Podpanel *other* je rezervisan za neke buduće verzije programa.

Pristup Bazi



Bazi podataka se pristupa iz glavnog panela ili, kao što je već spomenuto, u akvizicionom modu rada. Nakon što izabere objekat iz liste, korisnik može promeniti (ili popuniti ukoliko ne postoje) vrednosti sledećih polja:

- Ime grupe – Svaki objekat poseduje ime grupe kojoj pripada. Ime može biti proizvoljan niz znakova i služi kao kriterijum za grupisanje karakteristika objekata (površine i odnosa osa) u jednu jedinstvenu oblast;
- Tolerancija odnosa osa i površine - Definišu lokalnu oblast pripadnosti navedenih karakteristika datom objektu. Više ovih oblasti se spaja u jednu na osnovu informacija o imenu grupe;
- Orijentacija – Određuje da li algoritam treba da vrši procenu orijentacije objekta (odnosno njegov dela), ili je ona apriorno određena (na primer objekat je uvek položen);
- Regresivni model – Kao što je opisano može biti: linija, ravan ili kupa;
- Strategija hvata - Za dati 2D oblik propisuje strategiju hvata (lateralni, palmarni , ili prstima);
- Rotacija – Služi da bi se primitiv doveo u položaj u kome će referentne tačke obrazovati horizontalnu ili vertikalnu liniju (uvedeno kako bi se algoritam pretrage pojednostavio). Nema uticaja na rad programa;

- Promena referentnih tačaka – Definiše referentne tačke koje služe za procenu dela objekta od interesa. Algoritam automatski vrši njihovo poravnavanje (tako da budu ili horizontalne ili vertikalne), a na korisniku je da barem jednu tačku postavi na ivicu objekta (u suprotnom program vraća grešku i ignoriše unos).

Iz baze se mogu brisati postojeći objekti, a sve promene se vrše na njenoj kopiji (koja je prethodno učitana u memoriju) pa se po potrebi mogu odbaciti i vratiti na prethodno stanje.