STUDIA INFORMATICA 2012
Volume 33 Number 2B (106)

Grzegorz KOSZOWSKI[1], Michał KAWULOK[2]
Politechnika Śląska, Instytut Informatyki

VIRTUAL HAND MODELING FOR GESTURE RECOGNITION

Summary. In this paper we present a virtual hand tool for generating a database of artificial hand images for gesture recognition purposes. Our solution is based on 3D hand model with 23 degrees of freedom and skeleton animation. For every real hand image, the most similar artificial image is retrieved from the database, which allows hand pose be estimated. The initial results presented in the paper are encouraging and justify the further development of this method.

Keywords: gesture recognition, hand pose estimation, 3D modeling

MODELOWANIE LUDZKIEJ DŁONI NA POTRZEBY ROZPOZNAWANIA GESTÓW

Streszczenie. Celem prac przedstawionych w niniejszym artykule było opracowanie trójwymiarowego modelu ludzkiej dłoni na potrzeby rozpoznawania gestów. Model ten, posiadający 23 stopnie swobody i oparty na animacji szkieletowej, umożliwia generację bazy danych obrazów, pozwalającej na estymację parametrów ułożenia dłoni dla zadanego rzeczywistego obrazu. Uzyskane wyniki wstępne są satysfakcjonujące i wskazują na możliwość wykorzystania stworzonego modelu do rzeczywistych zastosowań.

Słowa kluczowe: rozpoznawanie gestów, estymacja ułożenia dłoni, grafika trójwymiarowa

[1] This work was supported by the European Union from the European Social Fund (grant agreement number: UDA-POKL.04.01.01-00-106/09).

[2] This work has been supported by the Polish Ministry of Science and Higher Education under research grant no. IP2011 023071 from the Science Budget 2012 – 2013.

1. Introduction

Since the beginnings of computing, many works have been focused on simplifying the ways, in which the computers can be operated by humans. Nowadays, human-computer interaction (HCI) constitutes an important field of computer science. Among many trends in simplifying the methods of interacting with the computers, recognizing hand gestures attracts an increasing attention [1, 2]. Gestures, understood as information conveyance throughout intentional movements of human body, play a substantial role in interpersonal communication, being one of its most natural ways.

Many gesture-based interfaces require additional equipment like magnetic gloves or multimedia sensors which retrieve location of characteristic hand landmarks [3, 4]. This makes them much more effective than the vision systems based exclusively on image analysis. However, following the opinion of Chaudhary et al. [3], due to additional cost and lack of comfort, such solutions are inapplicable in practice. It is worth to note that gesture recognition is successfully used in tablet interfaces or for augmented reality purposes. However, here the detection of the control points is hardware-based, and it is much simpler than automatic detection from images. The latter has become a very active research area of computer vision. The latest advance in gesture recognition and creating vision-based interfaces was introduced with presenting a Kinect sensor which acquires a depth map synchronized with standard 2D color image (it is also referred to as 2.5D). This simplifies the initial step of gesture recognition, i.e. skin segmentation and hand detection. However, estimating hand pose from 2D images remains a challenging and crucial problem. Altogether, the aforementioned facts justify the works focused on developing natural hand gesture recognition (HGR) interfaces based on vision information, acquired using a single camera.

A critical point in vision-based gesture recognition, regardless of whether 2D or 2.5D images are acquired, is to detect hand feature points and estimate the hand pose. Therefore, our work is focused on this aspect. Our general aim is to use a 3D hand model to generate artificial images of hand poses and elaborate the methods for comparing them with real images. In the work reported here, we present a Virtual Hand Tool (VHT) which is used for generating a database of artificial hand images that can be used for training the recognition systems. Here, we validate it with a naive approach, in which the gestures are classified using simple template matching, but we also outline how the tool will be used in future for real-world applications.

The paper is organized as follows. In Section 2 the state-of-the-art is presented with particular attention given to hand modeling. Our tool for automatic image generation is presented in Section 3, and in Section 4 it is explained how the hand pose can be estimated using the model. Section 5 holds the results of experimental validation, while in Section 6 we conclude the paper and present the directions of future work.

2. Background

2.1. Hand modeling

In general, hand pose recognition process is based on a database with “known” gestures descriptions. Depending on a detailed method, it can be an image, contour, distance map or a graph of points [1, 2, 5, 6]. Also, there are two main approaches to the model generation. Following the first one, it is necessary to gather the database from real examples and then learn the model using soft computing methods, while the second approach is to use a universal predefined model, based on which the gestures descriptions can be generated. In our approach we focused on the universal model defined in 3D space, and it is given more attention here.

There are two general groups of methods to store and deform a model, namely a) a skeleton with mesh, and b) cloud of points. Skeleton with mesh animation is used in 3D graphic and is called skeleton animation. Its hierarchical structure is composed of a root and bones connected to it. Each bone, except the root, has a parent (higher-level bone) and can have a child (lower level) and/or a sibling at the same level. When we move an individual bone, each bone from the lower level will be moved as well. This solution allows us to describe hand position with a description vector which contains joints. A joint is represented by three angles, i.e. rotations in 3D space. Following the cloud-of-points method, a hands is described using points in 3D space, called vertices. Such an approach is described in [7] and the model is built of 1,181 points. The main problem there is to divide those points in subdivisions and compute only a part of all 1,181 points. Each subdivision of the points is solved by vision estimator (VE) which is associated with possible controlling angles (like in skeleton animation). Computing all points with only three possible values of each angle needs around 10^9 tasks [6]. More information about that approach can be found in [6]. Overall, most of the existing works concerned with the gesture recognition is based on a 3D model using skeleton animation [2].

2.2. Related work

There have been many methods reported concerned with human hand pose estimation and gesture recognition. A simple approach is template matching of model-generated masks and real input image. Rosales [7] used a database of 9,000 records representing gestures with an assumption that they are exactly in front of camera and orthogonally positioned. Without those assumptions, the database reached 750,000 records. They reported to reach the accuracy of 70%, but using a database with 9,000 records and holding the strict assumption for hand position. Stenger [1] used 40 main templates for each gesture. In addition, every template was rotated 10 times from -10 to 10 degrees, which gives 400 templates per gesture. It is worth to note that in these both works virtual gloves were used to collect the model data.

There are also many methods, which do not use a hand model at all. They are based on feature points detection (like fingertips), contour-based recognition and hand networking. In a survey on gesture recognition [3], ten different works based on fingertips detection are reported. Most of them can operate only with open fingers. For closed fingers or fingers close to palm, the detection rate drops to only 10-20% [3]. The conclusion is that the fingertips-based methods cannot operate alone, but they can be used to support other approaches.

Contour analysis may also be used for hand pose estimation, either for direct comparison based on template matching or for feature extraction. Deng et al. proposed a method [8] for extracting contour features based on a set of the most relevant pixels. The pixels are selected by following the contour and analyzing the relative positions of subsequent contour pixels. The feature vector is formed of distances from the selected pixels to the palm center. Choi claim “the most important issue in field of the gesture recognition is the simplification of algorithm and the reduction of processing time” [9]. He proposes using morphological operation to implement system using the center points extracted from primitive elements by morphological shape decomposition.

Networking of a hand is also quite common. In [10] a self-growing and self-organized network is created in hand mask images. Usage of such solutions is especially effective with artificial neural networks (ANN) or other machine learning approaches. In [11] a pose recognition method which use geometric features is presented. It analyses the detected fingertips and base joints of all fingers. Classification is done by finding the minimal distance between equations from input image and model database. In [12] a histogram of mask is created around the mass center of hand. Classification is done by comparing histograms from real input to histograms database.

3. Virtual hand model

Among the most important problems concerned with HGR is high dimensionality of human hand [13]. It is constructed out of 27 bones, 8 of which are located at wrist. It gives 19 bones in palm and fingers. Joints are named according to their location in the hand. Metacarpophalangeal (MCP) join fingers to the palm, interphalangeal (IP) connect finger segments and carpometacarpal (CMC) they connect the metacarpal bones to the wrist. 9 IP joints in fingers have only one degree of freedom (DOF), 5 MCP have 2 DOF. From CMC joints, middle and index fingers are static, pinky and ring fingers have motion capability, but it is very limited and omitted in 3D modeling purposes. Thumb CMC in biomechanical studies have two non-orthogonal and non-intersecting rotation axes [14]. The wrist is usually represented as a 6-DOF joint. As a result, this gives a 28-DOF model.

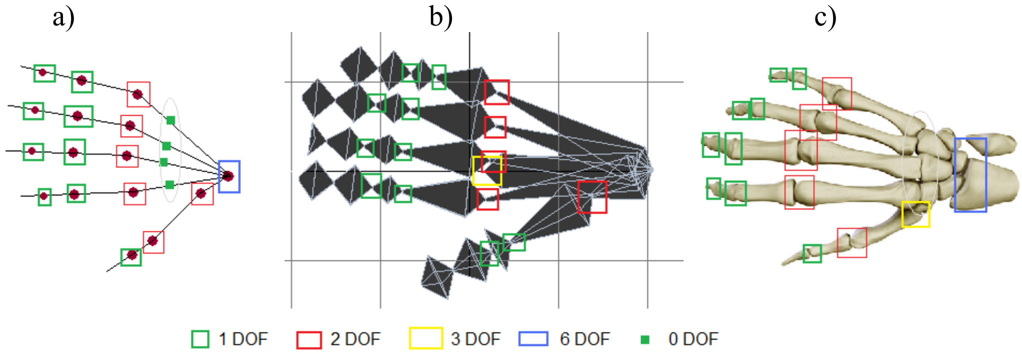

Fig. 1. Hand skeleton and DOF of the joints: a) commonly used, b) our model, c) real skeleton
Rys. 1. Szkielet dłoni i liczba stopni swobody (DOF) stawów: a) model najczęściej używany, b) zaproponowany model, c) rzeczywisty szkielet dłoni

In our solution we decided to simplify the model and reject some DOF in the following way:

- Thumb CMC is reduced to just 2 DOF joint, like it is the case in most of the posture recognition works.
- Thumb MPC is reduced from 2 DOF to 1 DOF. Of course, it is possible to use that rotation in our virtual hand tool but we assume it will not be used in our future works.
- Instead of modeling the wrist joint, we make it possible to rotate the hand around the central point of a sphere circumscribing the hand (the point is given 3 DOF).

As a result, this renders a 23-DOF model with possible usage of thumb MPC joint as 2 DOF which would expand our model by one DOF. After all, each gesture is described by a 23-dimensional vector. Each vector dimension represents one angle of an individual joint. Visualization of the skeleton is presented in Fig.1. Additional bones can be seen in our skeleton, which provide information about the fingertips and ground-truth hand center positions.

3.1. Virtual Hand Tool

DirectX is used as the graphic engine for skeleton animation and rendering hand images in our Virtual Hand Tool (VHT). The model is defined in DirectX files, which contain the polygons, their normal vectors, coordinates on the texture map and information about scene lighting. For image processing we use OpenCV library [15]. The default model is presented in Fig. 2. The most-right model is represented with non-uniform rational basis splines and the middle one is a standard polygon mesh. A hand image is generated based on a 23-dimensional vector representing all possible DOF of our model. Along with the image, its bitmap mask and a metadata layer are created. The metadata include 2D positions of the wrist, hand center and fingertips which are given from the model, and optionally a gesture identifier which can be used for indexing purposes.

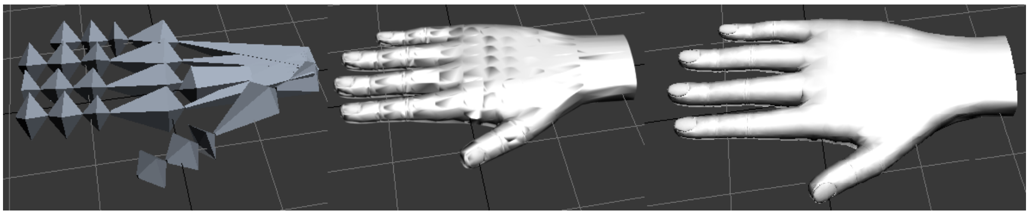

Fig. 2. A default model: skeleton (left) with two representations
Rys. 2. Domyślny model: szkielet (z lewej) i dwie reprezentacje

4. Pose recognition

VHT was used to generate a set of artificial reference images, based on which real hand images could be recognized. In the work reported here we used a naive approach for matching real images to those obtained using the model. First, the images are normalized to a fixed size, and it is provided that the hands are located in the same position. After that, the image binary masks are subject to template matching.

4.1. Reference database

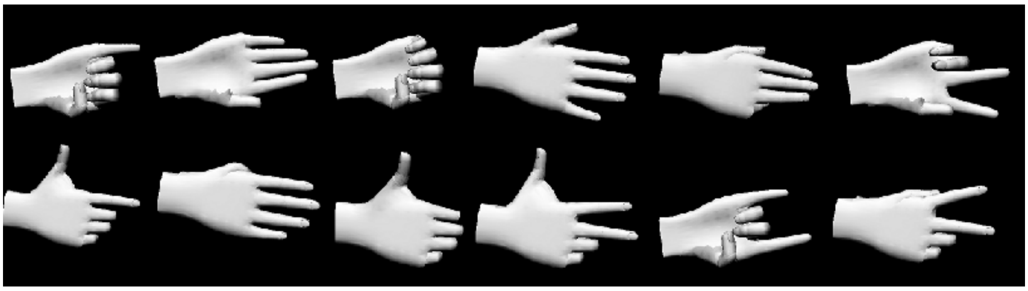

Fig. 3. Sample gestures from model database
Rys. 3. Przykładowe gesty z bazy danych modelu

Using VHT and default hand model presented in Section 3, we generated a sample reference database for 15 basic gestures. Each of them is represented using 2 to 36 samples, and the database contains 339 images. This is sufficient to validate our proof of concept, but in future definitely more images will have to be generated per every gesture. This will result in creating a large database, which will have to be efficiently indexed to make the method robust. Every image is associated with a gesture descriptor, based on which it can be easily retrieved from the database. Several reference images are presented in Fig. 3.

4.2. Image normalization

Hand images are subject to normalization based on palm center *(P)* and wrist point (*W*) location. When an image is obtained from the model, the feature points' locations are given. Otherwise, they have to be automatically extracted from input images, using skin region and wrist detection described in [16]. Palm center is considered as a mass center of the skin blob cropped at the wrist. The normalization process consists of a series of affine transforms, which are aimed at eliminating influence of rotation and scale variations. First, the image is rotated to horizontal orientation by an angle obtained from a vector *V*:

$$\vec{V} = \frac{\left(x_W - x_P, y_W - y_P\right)}{\left\|\left(x_W - x_P, y_W - y_P\right)\right\|}, \tag{1}$$

where x_P, y_P are coordinates of palm center pixel in the image plane and x_W, y_W are wrist coordinates. After that, the image is cropped and resized keeping two conditions: the wrist is located at the middle of the image left edge and the image is scaled so that it fills the destination image size of 100×100 pixels. The subsequent steps of the normalization procedure are presented in Fig. 4. First, the skin region and wrist are detected in the original image a) to determine the hand mask b), which is later rotated c), cropped and scaled d).

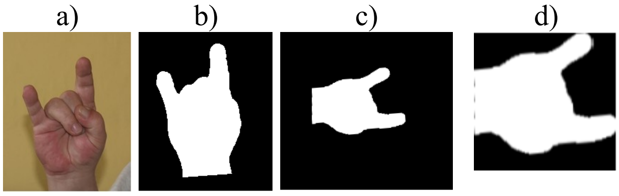

Fig. 4. Normalization operations
Rys. 4. Operacje normalizacji

4.3. Template matching

A real image, which is aimed to be recognized, is compared with all the samples in the database in order to determine the most similar artificial sample. As the gesture descriptor is known for every image in the database, also a hand pose can be estimated for the given image based on the closest match. Template matching after normalization is very simple and it is based on the difference between two masks:

$$R = \sum_{x,y} (T(x,y) - I(x,y))^2, \tag{2}$$

where *R* is the distance (i.e. the dissimilarity) between the masks, *T* is the real image mask and *I* is the mask obtained from the model. The classification is performed by finding the smallest *R* value.

5. Experimental validation

Our approach was validated based a database composed of 898 real hand images of 24 gestures. A smaller series holds 117 images of 10 gestures with hand positioned orthogonally plus/minus 10 degrees to camera. All experiments reported in this Section were performed using Intel Core 2 Duo P 8700 2,53 GHz processor and Radeon Mobile 4650 series graphic card.

Average time needed for hand mask generation, as well as for wrist detection and normalization is given in Table 1.

Table 1

Database generation times per image

Size	Type	Rendering time (ms)	Wrist detection time (ms)	Normalization time (ms)
332	real	X	2385	74
898	real	X	2004	77
337	model	25	2677	80

You can notice we had process wrist detection for a model. Fig. 5 presents an error understood as difference between wrist position in model and wrist position found in detection process. First histogram (a) shows error as Euclidean distance on two axes, second (b) only on X axis and third (c) on Y axis. Error value is calculated in percentage to the hand size. Model hand position is always the same. Hand is placed horizontally and orientated to the right. In Fig. 5 b) we can see that error is largest on X axis. Considering constant hand orientation and position we can deduce detected wrist position depends on hand pose. It justify usage of wrist recognition algorithm also for a model. The largest obtained error is for an image of hand with 2 pointing fingers (index and middle) and hidden thumb.

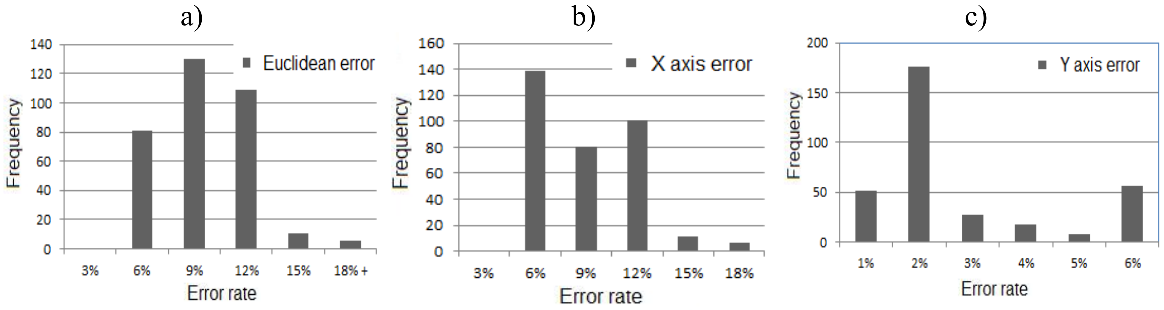

Fig. 5. Wrist detection error: a) Euclidean distance, b) only X axis, c) only Y axis
Fig. 5. Błąd detekcji nadgarstka: a) odległość euklidesowa, b) tylko oś X, c) tylko oś Y

Examples of template matching results are presented in Fig. 6. The upper series shows correctly recognized samples, while the lower demonstrate the matching errors. For every sample the real image (a) is shown with a hand mask (b), and the best match from the database is given in (c). The dissimilarity measure R is given below every sample (it is scaled by 10^{-8}). Overall, the accuracy for our main database of 898 images was 52,4 %.

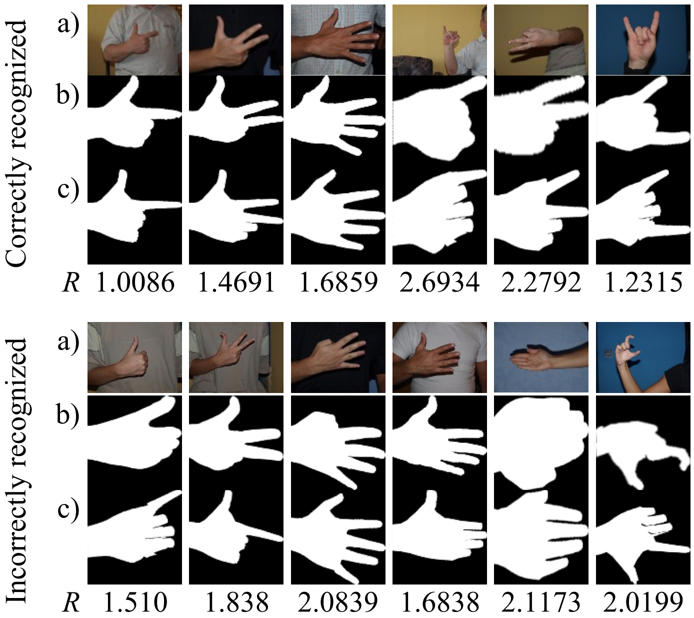
Correctly recognized
a)
b)
c)
R 1.0086 1.4691 1.6859 2.6934 2.2792 1.2315
Incorrectly recognized
a)
b)
c)
R 1.510 1.838 2.0839 1.6838 2.1173 2.0199

Fig. 6. Examples: a) of real images, b) hand masks, c) matched with model-generated masks
Fig. 6. Przykłady: a) rzeczywistych obrazów, b) masek dłoni, c) wraz z dopasowanymi maskami z modelu

Hand gesture recognition is a complex problem. In reality, people do not have identical hand shapes and, which is more important, they often present the same gesture in a slightly different manner. Such intra-class variations are illustrated in Fig. 7. Here, four real examples of the same gesture are shown, and it may be observed that they appear quite differently. This explains why every gesture has to be modeled using multiple reference images in a database to find the appropriate match.

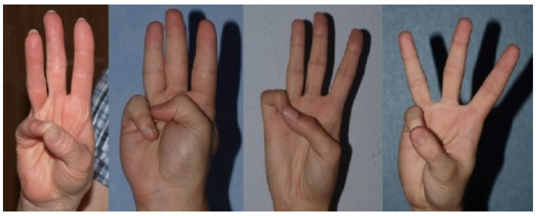

Fig. 7. Examples of the same gesture presented in a different manner
Rys. 7. Przykłady wykonania tego samego gestu na różne sposoby

Another essential problem is concerned with the wrist detection. It works well for the artificial model-based images, but worse with the images acquired in real environment. To investigate the influence of wrist detection inaccuracy we created a small database, which holds 117 images of 10 simple gestures without fingers crossed or overlapped. Moreover, we restricted that the hands are positioned orthogonally to the camera at a 10-degree tolerance. Here, the reference database contained 157 images generated from the model. Using the same classification methods as described in Section 4.3, the accuracy rate was 60 %. After that, we repeated the experiment for the same set of images, but the wrists locations were determined by human observers. This raised the accuracy up to 73%.

6. Conclusions and future work

In this paper we presented a tool based on a 3D hand model for gesture recognition purposes. The model defines a 23-dimensional parameter space, and every hand pose can be represented as a vector in this space. We validated the model-based approach using naive template matching method that compares hand masks extracted from real images with model-generated masks. We obtained accuracy of 73%, which may be considered as a high score, taking into account a fairly simple approach. This is a satisfactory result that allows for accepting this solution as a proof of concept.

Experimental validation revealed an extremely important aspect of hand pose estimation, which is the wrist localization, required for normalization purposes. For the automatic method we used, the detection errors decreased the recognition score to 60%. Therefore, in future we will have to focus on elaborating better methods for hand normalization and more accurate wrist detection.

Among other directions of future works, we want to extend the reference database substantially and introduce robust methods for indexing the hand poses based on features extracted from the images. This will determine a limited region from the entire parameter space that will have to be searched, which will make the matching procedure fast despite large number of samples in the database.

BIBLIOGRAPHY

1. Stenger B.: Template-Based Hand Pose Recognition Using Multiple Cues. Narayanan P.J. et al. (eds.): ACCV 2006, LNCS, Vol. 3852, Springer-Verlag Berlin Heidelberg 2006, p. 551÷560.
2. Erol S., Bebis G., Nicolescu M., Boyle R. D., Twombly X.: Vision-based hand pose estimation: A review. Computer Vision and Image Understanding, Vol. 108, 2007, p. 52÷73.
3. Chaudhary A., Raheja J. L., Das K., Raheja S.: Intelligent Approaches to interact with Machines using Hand Gesture Recognition in Natural way: A Survey. International Journal of Computer Science & Engineering Survey (IJCSES), Vol. 2, No. 1, 2011.
4. Causo A., Ueda E., Kurita Y., Matsumoto Y., Ogasawara T.: Model-based Hand Pose Estimation using Multiple Viewpoint Silhouette Images and Unscented Kalman Filter. Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, Technische Universität München, Munich, Germany 2008.
5. Chen J.-Y., Chang Y.-H.: A hand-pose recognition system using a combined classifier of shift distances and Fourier features. The 20th IPPR Conference on Computer Vision, Graphics and Image Processing (CVGIP), 2007.

6. Xu J., Wu Y., Katsaggelos S.: Part-based initialization for hand tracking. Proceedings of 2010 IEEE 17th International Conference on Image Processing, Hong Kong 2010.
7. Rosales R., Sclaroff S.: Combining Generative and Discriminative Models in a Framework for Articulated Pose Estimation. International Journal of Computer Vision, Vol. 67, 2006.
8. Deng L. Y., Lee D.-L., Keh H.-C., Liu Y.-J.: Shape context based matching for hand gesture recognition. IET International Conference on Frontier Computing. Theory, Technologies and Applications (CP568), Taichung, Taiwan 2010, p. 436÷444.
9. Choi J., Ko N., Ko D.: Morphological Gesture Recognition Algorithm. Proceeding of IEEE region 10th International Conference on Electrical and Electronic Technology, Coimbra, Portugal 2001, p. 291÷296.
10. Stergiopoulou E., Papamarkos N.: Hand gesture recognition using a neural network shape fitting technique. Engineering Applications of Artificial Intelligence, Vol. 22, Issue 8, 2009, p. 1141÷1158.
11. Bhuyan M.K., Neog D.R., Kar M.K.: Hand Pose Recognition Using Geometric Features. National Conference on Communications (NCC). Bangalore 2011.
12. Chakraborty P., Sarawgi P., Mehrotra A., Agarwal G., Pradhan R.: Hand Gesture Recognition: A Comparative Study. Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. 1, IMECS, Hong Kong 2008.
13. Xu Z., Zhu H.: Vision-based detection of dynamic gesture. International Conference on Test and Measurement, 2009, p. 223÷226.
14. Hollister A., Buford W.L., Myers L.M., Giurintano D.J., Novick A.: The axes of rotation of the thumb carpometacarpal joint. Journal of Orthopedic Research, Vol. 10 (3), 1992, p. 454÷460.
15. Bradski G.: The OpenCV Library. Dr. Dobb's J. of Software Tools, Vol. 25, No. 11, 2000, p. 120÷126.
16. Bernard S.: Ekstrakcja cech obrazu dłoni na potrzeby rozpoznawania gestów. Praca dyplomowa magisterska, Politechnika Śląska, Gliwice 2009.
17. Causo A., Matsuo M., Ueda E., Takemura K., Matsumoto Y., Takamatsu J., Ogasawara T.: Hand Pose Estimation using Voxel-based Individualized Hand Model. EEE/ASME International Conference on Advanced Intelligent Mechatronics, Suntec Convention and Exhibition Center, Singapore 2009.
18. Causo A., Ueda E., Kurita Y., Matsumoto Y., Ogasawara T.: Model-based Hand Pose Estimation using Multiple Viewpoint Silhouette Images and Unscented Kalman Filter. Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, Technische Universität München, Munich, Germany 2008.
19. Pickering C.A.: The search for a safer driver interface: a review of gesture recognition Human Machine Interface, IEEE Computing and Control Engineering, 2005, p. 34÷40.

Wpłynęło do Redakcji 22 stycznia 2012 r.

Omówienie

Rozpoznawanie gestów jest jedną z dynamicznie rozwijających się gałęzi wizji komputerowej. Wpływ na to mają nie tylko prace przy interfejsach człowiek-komputer, przeznaczonych dla osób niepełnosprawnych, ale również i coraz częstsze użycie modułów rozpoznawania gestów w rozwiązaniach komercyjnych, jak chociażby Wzbogacona Rzeczywistość (ang. *Augmented Reality*).

Proces rozpoznawania gestów często opiera się w literaturze na modelu 3D, przy użyciu którego generuję się referencyjną bazę danych. Wychodząc naprzeciw tej potrzebie, stworzyliśmy aplikację Virtual Hand Tool (VHT), która posłużyła nam w badaniach.

Przeprowadziliśmy pierwszy eksperyment przy użyciu VHT, generując referencyjną bazę danych dla modelu z 15 gestami, zawierającą łącznie 337 obrazów. Do tego celu użyliśmy stworzonego przez nas modelu dłoni. Drugą bazę danych uzyskaliśmy z prawdziwych zdjęć – 898 pozycji. Wyniki par obrazów uzyskaliśmy na podstawie najmniejszej odległości pomiędzy maskami 2 obrazów. Walidacje wyników przeprowadzono przez człowieka. Skuteczność detekcji, rozumiana jako procentowy udział poprawnie zakwalifikowanych obrazów z testowej bazy danych w stosunku do całkowitej ich ilości, wyniosła 52,4%. Dla bazy danych zawierającej proste gesty, bez krzyżujących się palców, zawierającej 117 obrazów 10 gestów, skuteczność naszego rozwiązania osiągnęła 60%. Po ręcznym wskazaniu nadgarstka, dla tej samej bazy danych, wzrosła do 73%. Zastosowana metoda okazała się bardzo wrażliwa na przesunięcia punktu nadgarstka. Pokazaliśmy również różnicę pomiędzy automatycznym wykryciem nadgarstka a jego umiejscowieniem w modelu. Wszystko to usprawiedliwia kierunek dalszych prac w celu lepszego wykrycia nadgarstka.

Wyniki okazały się wstępnie satysfakcjonujące, biorąc pod uwagę prostotę zastosowanej metody. W celu poprawienia wyników można zagęścić bazę danych kosztem czasu działania bądź też zdecydować się na metody bazujące na konturze. Tego typu metody, zaraz obok ulepszonej detekcji nadgarstka, będą kierunkiem naszych przyszłych badań.

Addresses

Grzegorz KOSZOWSKI: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16,
44-100 Gliwice, Poland, grzegorz.koszowski@polsl.pl.

Michał KAWULOK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16,
44-100 Gliwice, Poland, michal.kawulok@polsl.pl.