Journal of KONBiN 49(2019)
e-ISSN 2083-4608
DOI 10.2478/jok-2019-0011

Piotr GOLAŃSKI, Marek SZCZEKALA
Air Force Institute of Technology (Instytut Techniczny Wojsk Lotniczych)

THREE-DIMENSIONAL RECONSTRUCTION OF HAND USING STEREOSCOPIC IMAGES

Trójwymiarowa rekonstrukcja dłoni z wykorzystaniem obrazów stereoskopowych

***Abstract**: This article is devoted to works on using natural user interfaces (NUI) in computer support systems of aircraft service. The concept of such interfaces involves the usage in human-machine communication the same measures as in the communication between people, that is sound or gesture. In the case of gesture communication, it is indispensable to adopt methods related to computer vision algorithms. One of them is a three-dimensional reconstruction of objects based on processing techniques of a pair of two-dimensional images. The above method and the results of its application were presented to obtain a three-dimensional cloud of points describing the hand shape. The obtained software will constitute an element of gesture classifier based on the analysis of the spatial location of the acquired points of the cloud.*
Keywords: service of aircraft, NUI interfaces, computer vision, stereoscopy, image detection, OpenCV library

***Streszczenie:** Artykuł dotyczy prac nad wykorzystaniem naturalnych interfejsów użytkownika w komputerowych systemach wspomagania obsługi statków powietrznych. Koncepcja tego typu interfejsów zakłada wykorzystanie w komunikacji człowiek-komputer takich samych środków jak w komunikacji między ludźmi, a więc głosu lub gestu. W przypadku komunikacji za pomocą gestów konieczne jest zastosowanie metod związanych z algorytmami komputerowego widzenia. Jedną z nich jest trójwymiarowa rekonstrukcja obiektów oparta na technikach przetwarzania pary dwuwymiarowych obrazów. Przedstawiono tę metodę oraz wyniki jej zastosowania w celu uzyskania trójwymiarowej chmury punktów opisujących kształt dłoni. Uzyskane oprogramowanie będzie stanowić element klasyfikatora gestów opartego na analizie lokalizacji przestrzennej otrzymanych punktów chmury.*
Słowa kluczowe: obsługa statku powietrznego, interfejsy NUI, widzenie komputerowe, stereoskopia, rozpoznawanie obrazu, biblioteka OpenCV

1. Introduction

Using mobile systems of computer-based maintenance support fitted with Graphical User Interface (GUI) involves to constantly hold the device in hands. It is very problematic, especially during maintenance activities, during which both hands have to be used, e.g. to make regulation or use devices. In such a situation the operator is forced to put aside the device, losing the possibility to communicate with it.

In such situations, it is necessary to use computer software based on *wearable computers*, with which communication takes place using NUI (*Natural User Interface*). In interfaces of that type, communication is initiated by means of sound [5] or gesture.

Fig. 1. Using a head camera as an element of wearable computer

Using gestures for communication enables to resign from peripheral devices such as the keyboard or mouse and replace them with the operator's hands. It requires to determine the model reflecting the real hand in the computer system, and previously its recognition by the system.

Works on recognition of images, in particular gestures, are conducted throughout the world for decades and became applicable in smartphones. Such kinds of works are also carried out at ITWL [6, 7], and this article is a summary of their next step.

The stereo imaging method presented in this article enables to obtain a three-dimensional cloud of points and is based on processing techniques of a pair of images [2]. The article elaborates on the theoretical basis of the method and its

subsequent steps. Then, the results of the performance of software implementing this method are demonstrated. In the end, conclusions were formulated concerning the perspectives of its further usage.

2. Stereo imaging

Stereo imaging technique enables the formation in a digital form of a cloud of points reflecting the shape of the real object belonging to the physical world, on the basis of two pinhole cameras. Fig. 2 demonstrates an ideal model of stereo imaging.

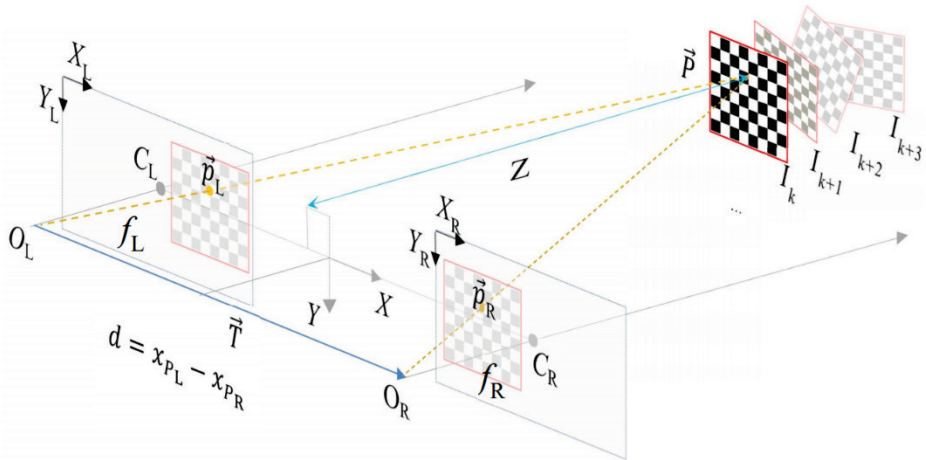


Fig. 2. Generation of stereo image

It is characterised by the following features:

1) camera images are coplanar,
2) optical axes of cameras are parallel to each other,
3) distance of optical axes of cameras $|\vec{T}|$ is known and constant,
4) focal lengths of cameras f are equal to each other, i.e.: $f = f_L = f_R$,
5) cameras are free from deformations of projected images,
6) principal points C have the same coordinates on their corresponding images i.e.: $c_x = c_{L_x} = c_{R_x}$,
7) projected images of cameras are row-aligned, i.e.: $y_L = y_R$,
8) real point $\vec{P}$ occurring in space has its equivalents $\vec{p}_{\mathrm{L}}$, $\vec{p}_{\mathrm{R}}$ on the images of corresponding cameras and their coordinates on the horizontal axis amount to $x_{p_{\mathrm{L}}}$, $x_{p_{\mathrm{R}}}$.

The above model allows for determining the relationship between the distance of point $\vec{P}$ from the stereo camera and the difference of locations x_{p_L}, x_{p_R} of their corresponding points $\vec{p}_L$, $\vec{p}_R$ on camera images.

From the similarity of triangles, the following can be written:

$$\frac{|\vec{T}|-d}{Z-f} = \frac{|\vec{T}|}{Z} \tag{1}$$

where d is a disparity – difference of locations between coordinates:

$$d = x_{p_L} - x_{p_R} \tag{2}$$

where: Z is depth – the distance of point $\vec{P}$ from the stereo camera.

Using relation (2) after the conversion of equation (1), we acquire the dependence of depth in disparity function $Z(d)$:

$$Z = \frac{f \cdot |\vec{T}|}{x_{p_L} - x_{p_R}} = \frac{f \cdot |\vec{T}|}{d} \tag{3}$$

Fig. 3 exhibits the dependence of disparity in depth function $d(Z)$.

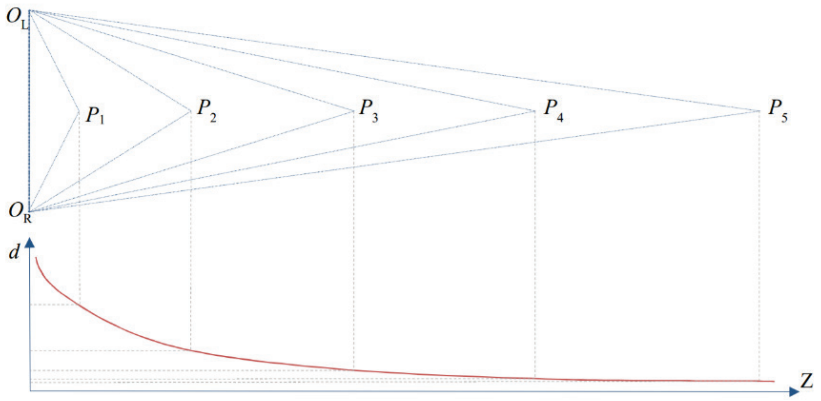


Fig. 3. Dependence of disparity from depth

As the figure implies, the disparity d is inversely proportional to the depth Z. In practical terms, this means that the resolution of depth is better the closer it is located to point P_k relative to the camera. This situation is very beneficial. After all, based on the human physical construction the distance Z is limited to the range

from 10 to approximately 70 cm (assuming the location of cameras is on the head as in fig. 1).

Since the stereo-processing model is not physically achievable, methods are applied, that enable one to mathematically convert the images obtained from real cameras to meet the Conditions 1 and 7 of the ideal model of stereo processing (fig. 2).

3. Reconstruction method of the spatial object

In the real *stereo imaging,* the majority of assumptions of the ideal model is not fulfilled. Especially, the cameras are not free from deformations of projected images. The image, projected onto the light-sensitive surface, is distorted (radial [3] and tangential distortions [4]). The sources of distortions are shapes and artefacts created during the production and precise assembly process of the lens.

The real location of point $\vec{P}$ is determined in four steps [2]:

1) undistortion,
2) rectification – transformation of orientations of a pair of cameras relative to each other,
3) correspondence – searching for the same features from images of both cameras,
4) reprojection – computing the depth for each of the detected feature (of the corresponding points from both cameras).

Undistortion of an image is a process, which consists in computing a matrix of internal parameters and distortion parameters by using specially designed calibration boards. Calibration of cameras involves applying a calibration board [11] with marked points with the known locations. In the case of a chessboard (fig. 2), such points are the internal corners of white and black squares. By registering the images of calibration boards from different angles, are computed four internal parameters of camera, (f_x, f_y, c_x, c_y), parameters of radial distortions k_i (their number is between 2 and 5 depending on the required precision), tangential distortions p_1, p_2 and a relative camera location defined by angles of rotation relative to the three axes and three-element translation vector in the function of each registered image I_k depicting a calibration board (fig. 2).

Using homographic transformations, that is the possibility to reconstruct one plane (calibration board) on the other plane, any number of equations can be achieved to compute all parameters as mentioned above necessary to remove camera distortions.

The objective of the next step of stereo processing, that is **rectification,** is to obtain a coplanar orientation and as precise the row equalization of images of two cameras as possible. It can be achieved by using Hartley's algorithm [8], but the results obtained by this method can be considerably distorted.

Better accuracy can be achieved using a Bouguet's algorithm [12]. This algorithm is implemented in OpenCV library [2] using internal parameters and distortion coefficients of two cameras and translation vectors $\vec{T}$ (fig. 2):

$$\vec{T} = \left[T_x, T_y, T_z\right]^T \tag{4}$$

and rotation matrix $\boldsymbol{R}$ defining their mutual location and orientation does the transformation to meet the Conditions 1 and 7 of the ideal model of stereo processing (Fig. 2).

To determine the third dimension – stereo depth (2), **correspondence** algorithms are applied. They search for their corresponding points form the left and the right camera. To this end, the OpenCV library employs two methods.

The first method, lying in *Block Matching* **(BM)** [10], is a fast method, but it enables to find only strongly matching points $\vec{p}_{\mathrm{L}}$, $\vec{p}_{\mathrm{R}}$ between rectified images of the left and the right camera. This algorithm is applicable when the image is full of numerous diversified patterns. In the case of scenes with a small number of characteristics (sky, dark empty room), the possibilities to determine depth for particular pixels are very limited.

The second method, called the *Semi-Global Block Matching* [9], is much slower because alignment takes place at subpixel level, using a Birchfield-Tomasi algorithm [1] employing global smoothness criterion in relation to calculated parameters of the scale of depth. Both methods used for finding matching pixels (points) in rectified images of both cameras use SAD (*Sum of Absolute Differences*) windows.

Regardless of the applied algorithm, as a result of their usage, for every point with the coordinates *x* and *y* we obtain a disparity *d*, from which at the **reprojection** step it is possible to determine a projection of a given point in three dimensions according to the following dependence:

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \\ \boldsymbol{W} \end{bmatrix} = \boldsymbol{Q} \cdot \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{d} \\ \boldsymbol{1} \end{bmatrix} \tag{5}$$

where:

X, Y, Z – not scalable three-dimensional coordinates,

W – coefficient of scale,

$\boldsymbol{Q}$ – matrix of mapping of disparities has depth with the following form:

$$\boldsymbol{Q} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & -\boldsymbol{c_x} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & -\boldsymbol{c_y} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{f} \\ \mathbf{0} & \mathbf{0} & -\frac{\mathbf{1}}{\boldsymbol{T_x}} & \mathbf{0} \end{bmatrix} \tag{6}$$

The above form of a matrix is true assuming that principal rays (optical axes, fig. 2) of both cameras cross in the infinity and c_x, c_y are coordinates of x and y of a principal point in the left image.

4. Results of hand's reconstruction

The theoretical fundamentals of the spatial object described in the previous point, based on its stereo imaging, became the basis for its programme implementation in C++ language in Visual Studio 2017. In the created software, many procedures were used relating to the image processing included in OpenCV library [2].

The processing used monochromatic images recorded by infrared cameras. The application of the infrared significantly decreased the time needed to prepare dynamically captured picture frames, due to the significant limitation of image background in relation to the foreground object (in our case: a hand). The change of foreground brightness is enabled by LED diodes. The diodes enlighten the space on the depth of 45 cm, which enables to obtain a homogeneous, dark grey background colour (background) (on the image).

The results of programme operation were presented in fig.4.

Fig. 4 shows a set of three columns of images. The left and right column exhibit unprocessed frames of raw images from the left and the right camera, with visible radial distortions. The middle column demonstrates the generated 3D images with the depth in the form of a grey scale in the range from 0 (averaged distance of camera lenses to the displayed pixel equals 0 on the image) to 255 (averaged distances of camera lens to the displayed pixel on the image is infinite, where: white colour = 0, black colour = 255.

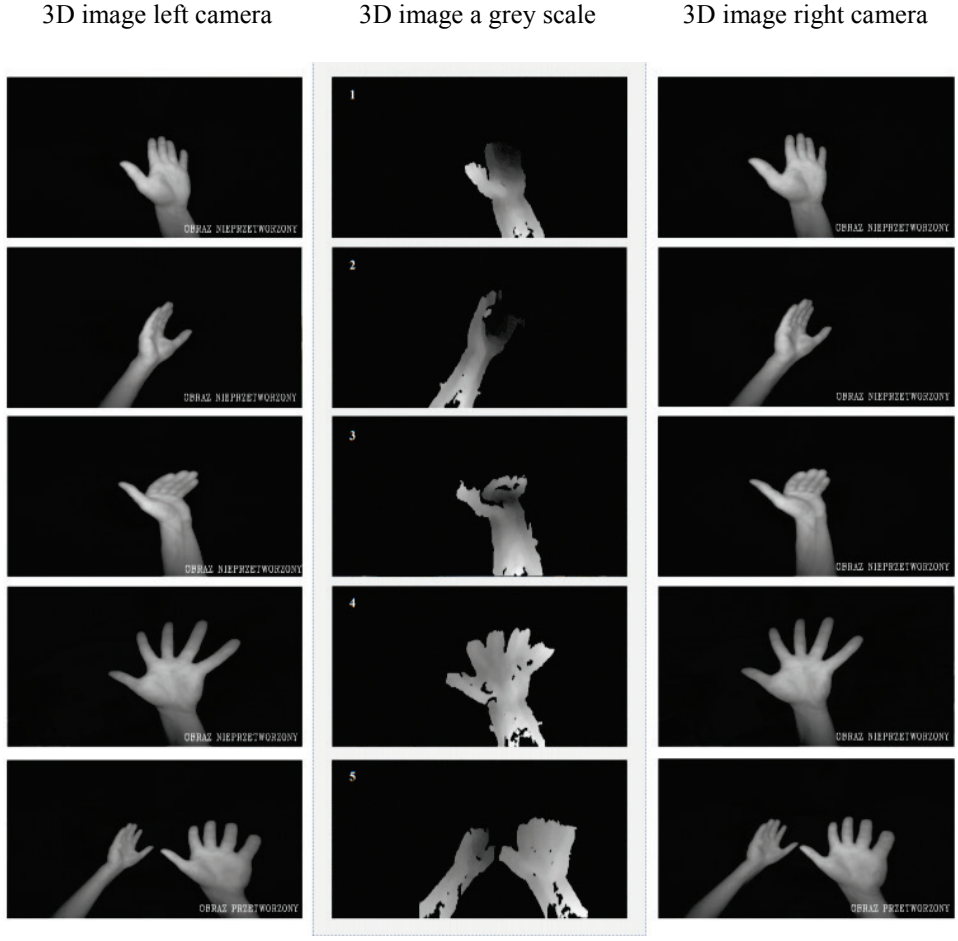


Fig. 4. Reconstruction of the spatial object (in the middle) based on images from the left and the right camera

Three-dimensional hand images presented in fig. 4. reflect the location of real hands relative to stereo images with precision. Hand elements (fingers), which can be observed on 3D images, are often connected to each other and not always clearly visible, which does not imply that it is impossible to determine its location. It should be noted that the depth on the presented 3D images contained in this article, is a standardised depth to up to 255 grey shades, where value 0 defines a black colour and 255 – white colour. The space enabling to detect the hand location is strictly limited, but for the use with VR (*virtual reality*) or AR (*augmented reality*) glasses does not appear to be a problem. Using the method of hand observation in the IR

band is a very good solution, due to the fact that the background is removed already by recording the image frame. The lack of a background on the images considerably simplifies the analysis of hand's contour and thus, speeds up the operation of the application.

5. Summary and Conclusions

This article described the method for stereo imaging enabling to obtain all detected features from the pair of two-dimensional images, allowing for the achievement of their three-dimensional location in the form of the cloud of points. Afterwards, the method was implemented with the application of the appropriate equipment, software and used for three-dimensional reconstruction of hand.

Taking into account all the obtained results included herein, it can be established that it is highly forward-looking method due to the fact that it enables the location and spatial orientation of hand, which allows for its reconstruction in the virtual world. The alignment of hands' location belonging to the real world with its virtual counterparts is thus a bridge enabling to create different interactions between these worlds.

The alignment of the obtained cloud of points reflecting the real hand with the virtual model of a hand will enable to relocate, move, turn, switch and touch with one's own hand the objects belonging to the virtual world. The next step of works can be an attempt to obtain the feeling of virtual objects by connecting them with the virtual counterparts.

Apart from this, obtaining a virtual hand model will allow for the construction of a gesture classifier based on it. Thus, the depicted method shall be used in human-machine communication algorithms as a software component of natural user interface (NUI) due to its high potential, which significantly exceels the possibilities of methods developed at ITWL in the previous years, which were earmarked only for gesture recognition based on the Chan-Ves algorithm or based on Fourier descriptors.

6. References

1. Birchfield S., Tomasi C.: Depth discontinuities by pixel-to-pixel stereo. Internatinal Journal of Computer Vision, Vol. 35, Iss. 3, 1999.
2. Bradski G., Kaehle A., Camera Models and Calibration & Projection and 3D Vision. Learnig OpenCV, 2008.

3. Brown D.C., Close-range camera calibration. Photogrammetric Engineering, 37, 1971.
4. Brown D.C., Decentric distortion of Lenses. Photogrammetric Engineering, 32, 1966.
5. Golański P., Szczekala M.: The voice interface implementation in the prototype of a mobile computer-aided aircraft technical support system. Zeszyty Naukowe Akademii Marynarki Wojennej - Scientific Journal of Polish Naval Academy, nr 2 (209)/2017.
6. Golański P., Szczekala M.: The analysis of the possibility of using Viola-Jones algorithm to recognize hand gesture in human-machine interaction. Aviation Advances & Maintenance, Vol. 40, Iss. 1, 2017.
7. Golański P., Szczekala M.: The structure of classifiers of hand gestures with the use of the active contour model and Fouriers descriptors. Aviation Advances & Maintenance, Vol. 41, Iss. 1, 2018.
8. Hartley R., Zisserman A.: Multiple View Geometry in Computer Vision. Cambridge University Press, 2006.
9. Hirschmuller H.: Stereo Processing by Semiglobal Matching and Mutual Information. Pattern Analysis and Machine Intelligence PAMI, 30, 2008.
10. Konolige K.: Small vision system: Hardware and implementation. Proceedings of the International Symposium on Robotics Research, Hayama 1997.
11. Zhang Z.: A fexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 2000.
12. http://www.vision.caltech.edu/bouguetj/ Complete Camera Calibration Toolbox for Matlab.

TRÓJWYMIAROWA REKONSTRUKCJA DŁONI Z WYKORZYSTANIEM OBRAZÓW STEREOSKOPOWYCH

1. Wstęp

Posługiwanie się mobilnymi systemami komputerowego wspomagania obsługi wyposażonymi w standardowy graficzny interfejs użytkownika GUI (*Graphical User Interface*) wymaga ciągłego utrzymywania urządzenia w dłoniach. Jest to bardzo problematyczne, szczególnie podczas wykonywania czynności obsługowych, w trakcie których należy zaangażować obydwie ręce, np. w celu wykonania regulacji czy też użycia narzędzi. Operator zmuszony jest wtedy odłożyć urządzenie, tracąc jednocześnie możliwość komunikacji z nim.

W takich sytuacjach konieczne jest wykorzystanie sprzętu komputerowego opartego na koncepcji komputerów do noszenia (*wearable computers*), z którymi komunikacja odbywa się z wykorzystaniem interfejsów NUI (*Natural User Interface*). W tego typu interfejsach komunikacja odbywa się głosem [5] lub gestem.

Rys. 1. Wykorzystanie nagłownej kamery jako elementu komputera do noszenia

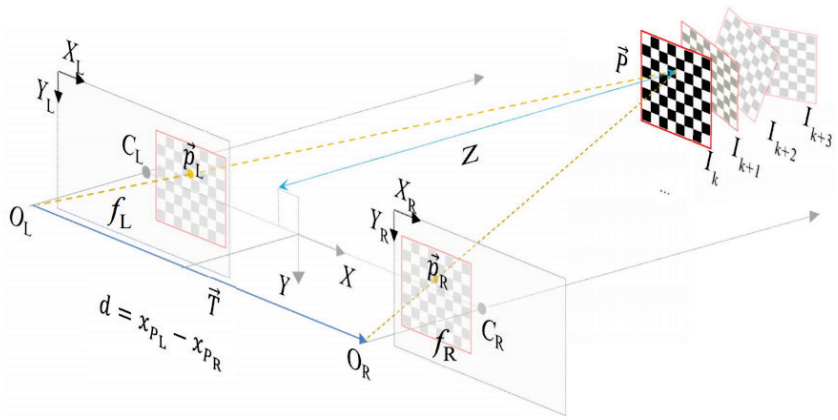
Wykorzystanie gestów do komunikacji pozwala na rezygnację z urządzeń peryferyjnych, takich jak klawiatura czy myszka, i zastąpienie ich rękami operatora. Wymaga to określenia modelu odwzorowującego rzeczywistą dłoń w systemie komputerowym, a wcześniej samego jej rozpoznania przez system.

Prace nad rozpoznawaniem obrazów, a w szczególności gestów, są prowadzone na świecie od dziesięcioleci i znalazły zastosowanie w tak popularnych obecnie urządzeniach, jakim są smartfony. Tego typu prace są prowadzone także w ITWL [6, 7], a niniejszy artykuł stanowi podsumowanie kolejnego ich etapu.

Przedstawiona w niniejszym artykule metoda obrazowania stereo umożliwia uzyskanie trójwymiarowej chmury punktów i jest oparta na technikach przetwarzania pary obrazów [2]. W artkule przedstawiono podstawy teoretyczne metody oraz kolejne jej etapy. Następnie zaprezentowano wyniki działania oprogramowania realizującego tę metodę. Na zakończenie sformułowano wnioski dotyczące perspektyw możliwości dalszego jej wykorzystania.

2. Zobrazowanie stereo

Technika obrazowania stereo umożliwia utworzenie w postaci cyfrowej chmury punktów odwzorowujących kształt rzeczywistego obiektu należącego do świata fizycznego, na podstawie dwóch dwuwymiarowych obrazów otrzymanych z dwóch kamer otworkowych. Na rys. 2 przedstawiono idealny model obrazowania stereo.



Rys. 2. Generacja obrazu stereo

Charakteryzuje się on tym, że:

1) obrazy kamer położone są współpłaszczyznowo,
2) osie optyczne kamer są do siebie równoległe,
3) odległość osi optycznych kamer $|\vec{T}|$ jest znana i niezmienna,
4) odległości ogniskowe kamer *f* są sobie równe, tj.: $f = f_L = f_R$,
5) kamery są wolne od deformacji rzutowanych obrazów,
6) punkty główne *C* mają takie same współrzędne na odpowiadających im obrazach tj.: $c_x = c_{L_x} = c_{R_x}$,
7) rzutowane obrazy kamer są wyrównane rzędowo, tj.: $y_L = y_R$,
8) rzeczywisty punkt $\vec{P}$ występujący w przestrzeni ma swoje odpowiedniki $\vec{p}_{\mathrm{L}}$, $\vec{p}_{\mathrm{R}}$ na obrazach odpowiednich kamer, a ich współrzędne na osi poziomej wynoszą $x_{p_{\mathrm{L}}}$, $x_{p_{\mathrm{R}}}$.

Powyższy model umożliwia wyznaczenie zależności pomiędzy odległością punktu $\vec{P}$ od kamery stereo a różnicą położeń $x_{p_{\mathrm{L}}}$, $x_{p_{\mathrm{R}}}$ odpowiadających im punktów $\vec{p}_{\mathrm{L}}$, $\vec{p}_{\mathrm{R}}$ na obrazach kamer.

Z podobieństwa trójkątów można zapisać:

$$\frac{|\vec{T}|-d}{Z-f} = \frac{|\vec{T}|}{Z} \tag{1}$$

gdzie *d* jest rozbieżnością – różnicą położeń pomiędzy współrzędnymi:

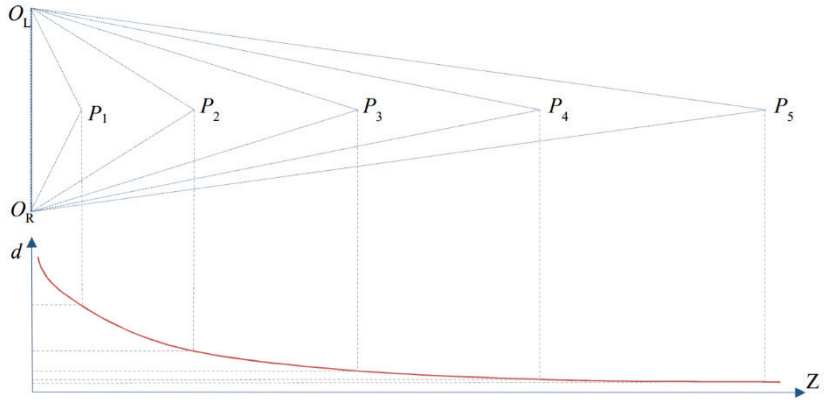
$$d = x_{p_{\mathrm{L}}} - x_{p_{\mathrm{R}}} \tag{2}$$

gdzie *Z* jest głębią – odległością punktu $\vec{P}$ od kamery stereo.

Wykorzystując zależność (2) po przekształceniu wzoru (1), otrzymujemy zależność głębokości w funkcji rozbieżności *Z*(*d*):

$$Z = \frac{f \cdot |\vec{T}|}{x_{p_{\mathrm{L}}} - x_{p_{\mathrm{R}}}} = \frac{f \cdot |\vec{T}|}{d} \tag{3}$$

Na rys. 3 przedstawiono zależność rozbieżności w funkcji głębi *d*(*Z*).



Rys. 3. Zależność rozbieżności od głębi

Jak wynika z rysunku, rozbieżność d jest odwrotnie proporcjonalna do głębi Z. W praktyce oznacza to, że rozdzielczość głębi jest tym lepsza, im bliżej kamery punkt P_k się znajduje. Jest to bardzo korzystne, gdyż z samego warunku budowy człowieka odległość Z jest ograniczona do zakresu od 10 do ok. 70 cm (zakładając umiejscowienie kamer na głowie jak na rys. 1).

Ponieważ przedstawiony model obrazowania stereo nie jest realizowalny fizycznie, stosowane są metody, które pozwalają przekształcić matematycznie obrazy otrzymane z rzeczywistych kamer, tak aby zostały spełnione warunki 1 i 7 idealnego modelu obrazowania stereo (rys. 2).

3. Metoda rekonstrukcji obiektu przestrzennego

W rzeczywistym obrazowaniu stereo (*stereo imaging*) nie jest spełniona większość założeń modelu idealnego. Przede wszystkim kamery nie są wolne od deformacji rzutowanych obrazów. Rzutowany na powierzchni światłoczułej obraz jest zniekształcony (zniekształcenia radialne [3] i tangensowe [4]). Źródłem zniekształceń są kształt, artefakty powstałe podczas produkcji oraz precyzja montażu soczewki.

Rzeczywiste położenie punktu $\vec{P}$ wyznaczane jest w czterech etapach [2]:

1) likwidacja zniekształceń obrazu (*undistortion*),
2) rektyfikacja (*rectification*) – transformacja położeń pary kamer względem siebie,
3) korespondencja (*correspondence*) – odnajdywanie tych samych cech w obrazach z obu kamer,

4) reprojekcja (*reprojection*) – obliczenie głębokości dla każdej z wykrytych cech (odpowiadających sobie punktów z obydwu kamer).

Likwidacja zniekształceń obrazu to proces polegający na wyliczeniu macierzy parametrów wewnętrznych oraz parametrów zniekształceń poprzez zastosowanie specjalnie przygotowanych plansz kalibracyjnych. Kalibracja kamer wymaga użycia planszy kalibracyjnej [11] z naniesionymi punktami o znanych położeniach. W przypadku planszy typu szachownica (rys. 2), tymi punktami są wewnętrzne narożniki białych i czarnych kwadratów. Rejestrując obrazy plansz kalibracyjnych pod różnymi kątami, wylicza się cztery parametry wewnętrzne kamery $\left(f_x, f_y, c_x, c_y\right)$, parametry zniekształceń radialnych k_i (ich liczba w zależności od wymaganej dokładności zawiera się pomiędzy 2 a 5) i tangensowych p_1, p_2 oraz względne położenie kamery określone kątami obrotu względem trzech osi oraz trójelementowym wektorem przesunięcia w funkcji każdego zarejestrowanego obrazu I_k przedstawiającego planszę kalibracyjną (rys. 2).

Wykorzystując przekształcenia homograficzne, tj. możliwość odwzorowania jednej płaszczyzny (plansza kalibracyjna) na inną płaszczyznę (płaszczyzna obrazująca) możemy otrzymać dowolną liczbę równań, tak aby obliczyć wszystkie wymienione powyżej parametry konieczne do usunięcia zniekształceń kamer.

Celem następnego etapu przetwarzania w kierunku uzyskania idealnego zobrazowania stereo, czyli **rektyfikacji**, jest uzyskanie współpłaszczyznowej orientacji i jak najdokładniejszego rzędowego wyrównania obrazów dwóch kamer. Można to osiągnąć, stosując algorytm Hartleya [8], jednak uzyskane tą metodą wyniki mogą być znacznie zniekształcone. Większą dokładność można uzyskać, stosując algorytm Bougueta [12]. Algorytm ten zaimplementowany w bibliotece OpenCV [2], wykorzystując wewnętrzne parametry i współczynniki zniekształceń dwóch kamer oraz wektorów translacji $\vec{T}$ (rys. 2):

$$\vec{T} = \left[T_x, T_y, T_z\right]^T \tag{4}$$

oraz macierzy rotacji $\boldsymbol{R}$ wiążących ich wzajemne położenie i orientację, dokonuje transformacji, tak aby spełnić warunki 1 i 7 idealnego modelu obrazowania stereo (rys. 2).

Do wyznaczenia trzeciego wymiaru – głębi stereo (2) stosowane są algorytmy **korespondencji** – wyszukujące odpowiadające sobie punkty z lewej i prawej kamery. W tym celu w bibliotece OpenCV wykorzystywane są dwie metody.

Pierwsza metoda, polegająca na dopasowywaniu bloków BM (*Block Matching*) [10], jest metodą szybką, jednak umożliwia znajdowanie tylko silnie

pasujących punktów $\vec{p}_{\mathrm{L}}$, $\vec{p}_{\mathrm{R}}$ ze zrektyfikowanych obrazów lewej i prawej kamery. Algorytm ten sprawdza się w przypadku występowania na obrazie licznych zróżnicowanych wzorów. W przypadku scen z małą ilością wyróżników (niebo, ciemne puste pomieszczenie) możliwości wyznaczania głębi dla poszczególnych pikseli są bardzo ograniczone.

Druga metoda, nazywana semiglobalnym dopasowywaniem bloków SGBM (*Semi-Global Block Matching*) [9], jest znacznie wolniejsza, ponieważ dopasowanie odbywa się na poziomie subpikselowym, wykorzystującym algorytmy Birchfielda-Tomasiego [1] przy zastosowaniu globalnego kryterium gładkości w odniesieniu do wyliczonych parametrów skali głębi. Obydwie metody do wyszukiwania pasujących do siebie pikseli (punktów) w zrektyfikowanych obrazach obu kamer wykorzystują okna sum różnic bezwzględnych SAD (*Sum of Absolute Differences*).

Niezależnie od zastosowanego algorytmu, w wyniku ich stosowania dla każdego punktu o współrzędnych *x* i *y* otrzymujemy rozbieżność *d*, z której na etapie **reprojekcji** można wyznaczyć rzut danego punktu w trzy wymiary według poniższej zależności:

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \\ \boldsymbol{Z} \\ \boldsymbol{W} \end{bmatrix} = \boldsymbol{Q} \cdot \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{d} \\ \mathbf{1} \end{bmatrix} \tag{5}$$

gdzie:
X, Y, Z – nieprzeskalowane trójwymiarowe współrzędne,
W – współczynnik skali,
$\boldsymbol{Q}$ – macierz odwzorowań rozbieżności na głębię o postaci:

$$\boldsymbol{Q} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & -\boldsymbol{c_x} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & -\boldsymbol{c_y} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{f} \\ \mathbf{0} & \mathbf{0} & -\frac{\mathbf{1}}{\boldsymbol{T_x}} & \mathbf{0} \end{bmatrix} \tag{6}$$

Powyższa postać macierzy jest prawdziwa przy założeniu, że promienie główne (osie optyczne, rys. 2) obydwu kamer przecinają się w nieskończoności natomiast c_x, c_y są współrzędnymi *x* i *y* punktu głównego w lewym obrazie.

4. Wyniki rekonstrukcji dłoni

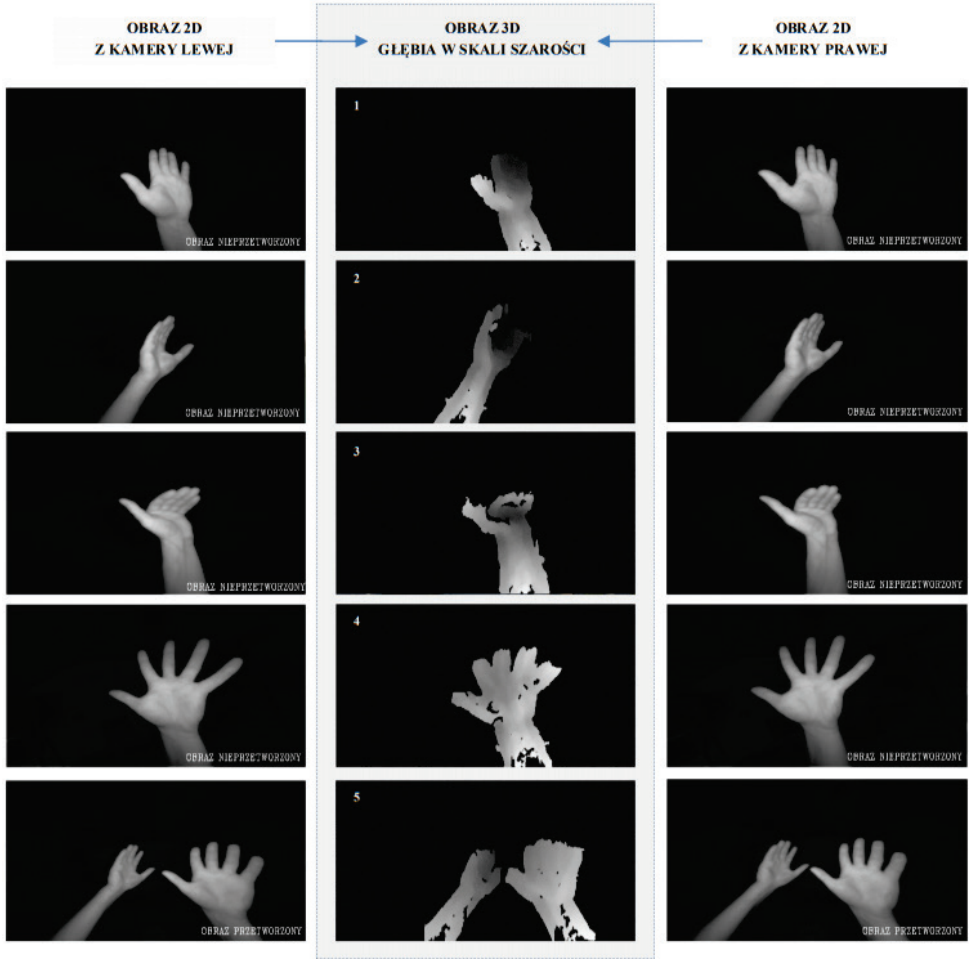
Przedstawione w poprzednim punkcie teoretyczne podstawy rekonstrukcji obiektu przestrzennego na podstawie jego zobrazowania stereo stały się podstawą do ich implementacji programowej w języku C++ w środowisku Visual Studio 2017. W powstałym oprogramowaniu wykorzystano wiele procedur związanych z obróbką obrazu zawartych w bibliotece OpenCV [2].

Do obróbki wykorzystano obrazy monochromatyczne rejestrowane przez kamery w paśmie podczerwieni. Zastosowanie podczerwieni w znaczący sposób skróciło czas konieczny do przygotowania dynamicznie przechwytywanych ramek zdjęciowych, ze względu na bardzo wyraźne ograniczenie dominacji tła obrazu w stosunku do obiektu występującego na pierwszym planie (w naszym przypadku dłoni). Zmianę jaskrawości planu pierwszego w stosunku do pozostałych umożliwiają diody podświetlające LED. Diody doświetlają przestrzeń na głębokość do 45 cm, co umożliwia uzyskanie prawie jednorodnego, ciemnoszarego koloru tła (drugi plan) na obrazie.

Wyniki działania programu przedstawiono na rys. 4.

Rys. 4 stanowi zestaw trzech kolumn zdjęć. Lewa i prawa kolumna przedstawiają nieprzetworzone ramki obrazów surowych z lewej i prawej kamery, z widocznymi zniekształceniami radialnymi. Środkowa kolumna przedstawia wygenerowane wynikowe obrazy 3D z głębią w postaci skali szarości w zakresie od 0 (uśredniona odległość obiektywów kamer do wyświetlanego piksela na zdjęciu jest równa 0) do 255 (uśredniona odległość obiektywów kamer do wyświetlanego piksela na zdjęciu jest równa nieskończoność) gdzie: kolor biały = 0, kolor czarny = 255.

Przedstawione na rys. 4 trójwymiarowe obrazy dłoni w bardzo precyzyjny sposób odzwierciedlają położenie rzeczywistych dłoni względem kamer stereo. Elementy dłoni (palce), które można obserwować na obrazach 3D są często ze sobą połączone i nie zawsze dobrze widoczne, co nie oznacza, że ich położenie jest niemożliwe do wyznaczenia. Należy bowiem pamiętać, że głębokość na przedstawianych w niniejszej pracy obrazach 3D jest głębokością znormalizowaną do jedynie 255 odcieni szarości, gdzie wartość 0 określa kolor czarny, a 255 kolor biały. Przestrzeń umożliwiająca wykrywanie położenia dłoni jest mocno ograniczona, jednak do zastosowań z okularami VR (*virtual reality*) lub AR (*augmented reality*) wydaje się nie stanowić przeszkody. Zastosowanie metody obserwacji dłoni w paśmie IR jest bardzo dobrym rozwiązaniem, ponieważ już na wysokości rejestracji klatki zdjęciowej usuwane jest tło. Brak drugiego planu na zdjęciu w znacznym stopniu upraszcza analizę konturu dłoni, a tym samym przyspiesza pracę aplikacji.



Rys. 4. Rekonstrukcja obiektu przestrzennego (w środku) na podstawie obrazów z lewej i prawej kamery

5. Podsumowanie i wnioski

W niniejszym artykule przedstawiono sposób uzyskania obrazowania stereo, umożliwiający otrzymanie wszystkich rozpoznanych cech z pary obrazów dwuwymiarowych, pozwalając na uzyskanie ich trójwymiarowej lokalizacji w postaci chmury punktów. Następnie metoda została zaimplementowana

z zastosowaniem odpowiedniego sprzętu oraz oprogramowania i zastosowana do trójwymiarowej rekonstrukcji dłoni.

Biorąc pod uwagę uzyskane wyniki, można stwierdzić, że jest to metoda wyjątkowo perspektywiczna, ponieważ umożliwia lokalizację i przestrzenną orientację dłoni, co pozwala na jej odtworzenia w świecie wirtualnym. Powiązanie położeń dłoni należących do świata rzeczywistego z ich wirtualnymi odpowiednikami, jest bowiem pomostem umożliwiającym kreowanie różnorodnych interakcji pomiędzy tymi światami.

Powiązanie uzyskanej chmury punktów odzwierciedlającej rzeczywistą dłoń z wirtualnym modelem dłoni umożliwi przenoszenie, przesuwanie, obracanie, przełączanie, dotykanie własną dłonią przedmiotów należących do świata wirtualnego. Kolejnym etapem prac może być próba uzyskania czucia przedmiotów wirtualnych poprzez sprzęganie ich z ich rzeczywistymi odpowiednikami.

Poza tym uzyskanie wirtualnego modelu dłoni pozwoli na zbudowanie w oparciu o niego klasyfikatora gestu. Dlatego przedstawiona metoda powinna być wykorzystywana w algorytmach komunikacji człowiek-komputer jako komponent oprogramowania naturalnego interfejsu użytkownika NUI ze względu na jej wielki potencjał, który znacznie przewyższa możliwości opracowanych w poprzednich latach w ITWL metod, przeznaczonych jedynie do rozpoznawania gestów na bazie algorytmu Chana-Vese [6] lub w oparciu o deskryptory Fouriera [7].

6. Literatura

1. Birchfield S., Tomasi C.: Depth discontinuities by pixel-to-pixel stereo. Internatinal Journal of Computer Vision, Vol. 35, Iss. 3, 1999.
2. Bradski G., Kaehle A., Camera Models and Calibration & Projection and 3D Vision. Learnig OpenCV, 2008.
3. Brown D.C., Close-range camera calibration. Photogrammetric Engineering, 37, 1971.
4. Brown D.C., Decentric distortion of Lenses. Photogrammetric Engineering, 32, 1966.
5. Golański P., Szczekala M.: The voice interface implementation in the prototype of a mobile computer-aided aircraft technical support system. Zeszyty Naukowe Akademii Marynarki Wojennej - Scientific Journal of Polish Naval Academy, nr 2 (209)/2017.
6. Golański P., Szczekala M.: The analysis of the possibility of using Viola-Jones algorithm to recognize hand gesture in human-machine interaction. Aviation Advances & Maintenance, Vol. 40, Iss. 1, 2017.
7. Golański P., Szczekala M.: The structure of classifiers of hand gestures with the use of the active contour model and Fouriers descriptors. Aviation Advances & Maintenance, Vol. 41, Iss. 1, 2018.

8. Hartley R., Zisserman A.: Multiple View Geometry in Computer Vision. Cambridge University Press, 2006.
9. Hirschmuller H.: Stereo Processing by Semiglobal Matching and Mutual Information. Pattern Analysis and Machine Intelligence PAMI, 30, 2008.
10. Konolige K.: Small vision system: Hardware and implementation. Proceedings of the International Symposium on Robotics Research, Hayama 1997.
11. Zhang Z.: A fexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 2000.
12. http://www.vision.caltech.edu/bouguetj/ Complete Camera Calibration Toolbox for Matlab.