Face Profile View Retrieval Using Time of Flight Camera Image Analysis

Piotr Bratoszewski(✉) and Andrzej Czyżewski

Multimedia Systems Department, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland
{bratoszewski,andcz}@sound.eti.pg.gda.pl

Abstract. Method for profile view retrieving of the human face is presented. The depth data from the 3D camera is taken as an input. The preprocessing is, besides of standard filtration, extended by the process of filling of the holes which are present in depth data. The keypoints, defined as the nose tip and the chin are detected in user's face and tracked. The Kalman filtering is applied to smooth the coordinates of those points which can vary with each frame because of the subject's movement in front of the camera. Knowing the locations of keypoints and having the depth data the contour of the user's face a profile retrieval is attempted. Further filtering and modifications are introduced to the profile view in order to enhance its representation. Data processing enhancements allow emphasizing minima and maxima in the contour signals leading to discrimination of the face profiles and enable robust facial landmarks tracking.

Keywords: Depth image · Signal processing · Profile view · Keypoints tracking

1 Introduction

The profile view of human face has multiple applications and robust methods of its retrieving are regarded as highly necessary. Historically, one of the first serious applications of the profile view were the mug shot photographs of the persons after being arrested and they are dated back to 1844. Mug shot photography consists of frontal and profile photography of a person suspected of a crime and the purpose of those is to create a record and identify criminals. Hence, the profile view is considered to contain enough biometric data to be successfully used in face recognition domain.

Furthermore, the profile view is known to be used in lip-reading applications. The Audio-Visual Speech Recognition Systems (AVSR) use the visual data to improve the accuracy of AVSR system in the noisy environments. To name a few applications: systems of emotion recognition or facial actions, methods for face modeling and texturizing and more.

However, authors of this work are focused on the way of retrieving of the profile view rather than on the variety of applications. Retrieval of the profile view can be achieved by using either one RGB camera placed on the side of the user or two cameras in front of the subject – in the stereo configuration. The first solution is considered to be both inconvenient (e.g. user of the mobile computer requires the eye contact with the

M. Kryszkiewicz et al. (Eds.): PReMI 2015, LNCS 9124, pp. 159–168, 2015.
DOI: 10.1007/978-3-319-19941-2_16

platform in which the camera is built into) and resource-intensive as it needs to employ additional background subtraction techniques which are susceptible to light conditions and deal poorly with non-static backgrounds. The stereo configuration of RGB cameras increases the overall cost of the system as for good depth estimation high quality sensors must be used and additional processing must be applied in order to extract the depth information from the image. Therefore, authors use the newer technology, namely the Time of Flight (TOF) 3D camera which provides the depth data on its output using integrated dedicated sensors and microprocessors for the depth calculation. TOF cameras are less sensitive to lighting conditions, hence the background subtraction is based on a proper thresholding rather than on the modeling of the background, making it possible to extract the foreground robustly, also when the background is non-static. One of main disadvantages of TOF sensors is low spatial resolution, however it is improving over time.Utilizing depth sensors for the task of profile view retrieval is more convenient than RGB sensor, as the TOF camera acquires the depth information of user's head directly and can be mounted for example in a laptop or any mobile platform in order to gather data while the user is simply using the device. The process of gathering and analyzing of such data will be presented in a more detailed way in next sections.

The paper is organized as follows: in the next section authors present their related work to the subject of profile view retrieving and usage. Section 3 describes the methods applied in this work. Section 4 presents the final results of conducted experiments and Sect. 5 concludes the paper.

2 Related Work

The significant interest in automatic face profile view retrieval has been reported in literature. There are many studies concerning supporting of the Automatic Speech Recognition using visual signal in noisy environments. Kumar et al. proposed usage of the profile view for lip reading [1]. Their study involves usage of RGB camera standing on the side of the subject and the blue color background for easy foreground extraction using color thresholding. Authors of the mentioned paper use the dictionary consisting of 150 words and their system was speaker-dependent. The WER gain from using the profile view geometrical features was 39.6 % in environment contaminated by noise of SNR equal to -10 dB. More researches concerning profile view in field of AVSR can be found in the literature. Worth noticing are the works of Lucey and Potamianos [2], Pass et al. [3] and Navarathna et al. [4].

Dalka et al., present another approach to AVSR system where the visual features are based on Active Appereance Models (AAM) [5]. Geometrical features of lips image acquired by the RGB camera pointing the subject are proposed. The database used by authors in their work is described in detail in the paper by Kunka et al. [6]. The language corpus adopted in this database is based on the thorough studies of the natural English characteristics by Czyzewski et al. who prepared the language sample that reflects the vowel and consonant frequencies in natural speech for the purposes of training the AVSR systems [7]. The result of the study by Dalka et al. proves that the lip contour detection algorithm is reliable and accurate for visual speech recognition tasks.

Zhou and Bhanu proposed human recognition system based on face profile views in video [8]. Their approach uses high resolution face profile images constructed from low resolution videos from a side view camera. Authors used their own method for feature extraction known as curvature estimation and the classifier applied to match profile views is based on the dynamic time warping method. On the average, more than 70 % of persons were correctly recognized by usage of face profile.

A large study on profile view keypoints detection and tracking is presented by Pantic and Patras [9]. Authors studied the role of facial expressions that provide a number of social signals while people communicate to each other. They proposed a system for automatic facial actions recognition based on 15 keypoints found in the profile view image from RGB camera. The average recognition rate of the 27 facial actions tested in the work was equal to 86.6 %.

Major of studies to date are concerned on profile view retrieved by the RGB cameras. Therefore, authors of this work found it necessary to use the new type of data provided by depth imaging cameras in order to retrieve the facial profile view of the subjects.

3 Methods

In this paper the workflow depicted in Fig. 1 is used for retrieving the profile view of the user's face. After the frame acquisition, the preprocessing is needed to smooth the depth map of the user's face and to filter out the artifacts present in the image. Having the preprocessed frame, the nose and chin detection are performed in order to fit the face bisection line in the next stage. Knowing the line which passes through the center of the face, the profile view based on the depth data can be retrieved and the facial landmarks can be tracked. A more detailed description of all modules is presented in further subparagraphs.

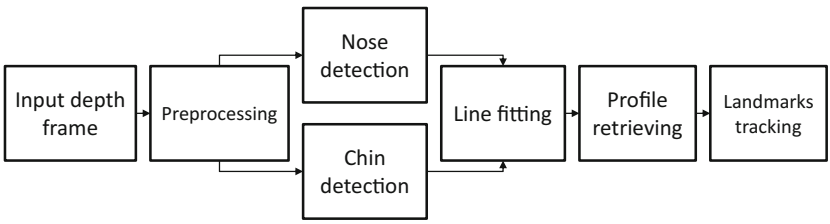

Fig. 1. Workflow for face profile retrieving method

3.1 Image Acquisition

In this work the Softkinetic DepthSense 325 Time-of-Flight camera was used. This camera enables imaging of the depth of the scene. Depth frames are acquired at 60 frames per second with resolution of 320 per 240 pixels, whereas data are transferred to PC using the USB interface. For the application of accurate face profile retrieving the acquired depth map is too noisy, hence, the preprocessing had to be applied, as is discussed in the following paragraph.

3.2 Preprocessing

The preprocessing is focused on denoising of the input data without degrading the useful signal. The preprocessing, which stages are depicted in Fig. 2, is split into two main subprocesses – the first one is a classical filtering using median and bilateral filters [10, 11], the second one is a so-called hole filling process [12, 13]. As it can be seen in Fig. 6. On the right, the depth signal is accompanied with large amount of noise. Therefore it must be subjected to an appropriate filtration. Two filtration methods were chosen – median and bilateral ones. The kernel of median filter used is equal to 5, so the examined neighborhood is of size 5 × 5 pixels with the purpose to remove the pepper noise from the signal. A bilateral filter is chosen to perform a more thorough filtration as bilateral filters are known to be able to denoise the signal preserving the information concerning the shape of photographed image edges. The weights of the bilateral filter are calculated as in Eq. 1 which is derived from the work of Barash [14]:

$$w(i,j,k,l) = e^{\left(-\frac{(i-k)^2+(j-l)^2}{2\sigma_d^2} - \frac{\|I(i,j)-I(k,l)\|^2}{2\sigma_r^2}\right)}, \tag{1}$$

where i, j are pixel locations being filtered, k, l are the neighboring pixels I denotes the pixel intensity and σ_d and σ_r represent smoothing parameters. In experiments described later on, the neighborhood size was set equal to 9 × 9 pixels and the sigma color σ_r was set to 20, whereas the sigma space σ_d was set to 8. The values of constants and neighborhood sizes in median and bilateral filtration methods were chosen empirically for optimal balance between noise removal and information preservation. Values had to be chosen with respect to moderately low depth image resolution of 320 per 240 pixels.

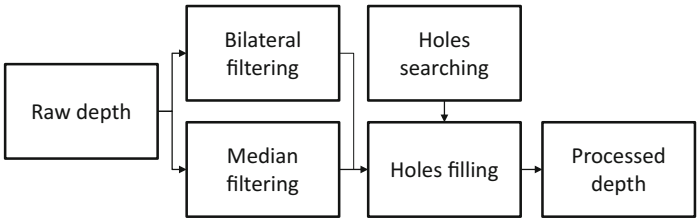

Fig. 2. Process of filtering and holes-filling of depth data

The hole filling (morphological processing) is a necessary stage as image data holes is a common issue of depth imaging cameras based on near-infrared light. It is mainly caused by the highly reflective surfaces like polished metals or the glass, hence, in our application it mainly occurs when the user wears glasses. Another source of these artifacts is fast movement, as the TOF cameras do not handle it very well. The depth holes phenomena is depicted in Fig. 3 as well as the result of holes filling process.

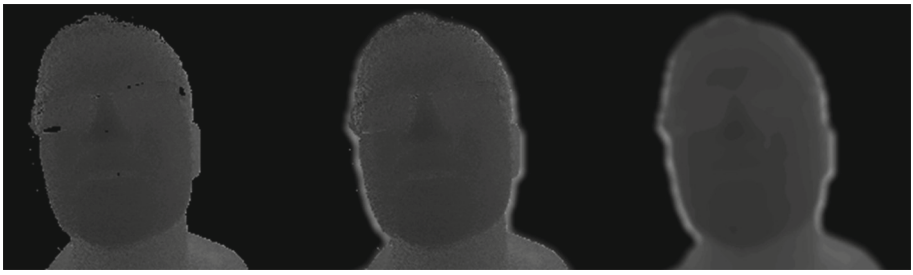

Fig. 3. Holes-filling process; left: raw depth frame, center: face with filled holes, right: filtered data used for hole-filling

The authors approach to the process of holes filling consists of following steps:

- Creating a binary mask of the user's silhouette
- Skimming through every pixel of depth map inside of the binary mask to find irrelevant pixels with values equal to 0 or 1 (holes can be seen in Fig. 3 on the left)
- Values of these pixels are replaced with the values of strongly filtered depth map which does not contain any hole (depicted in Fig. 3 on the right)

This process results in a depth image of the user without image holes (Fig. 3. in the center) with accurate depth data. Only the holes locations are filled with approximation values from the filtered depth data.

3.3 Keypoints Detection and Tracking

Keypoints are defined in this paper as points representing the tip of the nose and the chin. Those points are found in template matching process, using the artificially prepared models of the nose and the chin which are depicted in Fig. 4. In order to create those models, the patches of nose and chin regions were extracted from the depth images of the subjects. Furthermore, they were artificially modified in order to enhance the contrast between the close and far pixels using level based transform which lowers the intensity of dark pixels and increases the value of bright pixels. The process of localization is illustrated in Fig. 5. The template matching method utilizes the normalized cross correlation coefficient in order to find the similar part to template in the image. This method is described in more detail in work of Briechle and Hanebeck [15] and it adopts the following equation to calculate the values of the coefficient (Eq. 2):

$$R(x,y) = \frac{\sum_{x',y'}(T'(x',y') * I'(x+x',y+y'))}{\sqrt{\sum_{x',y'}(T'(x',y')^2 * \sum_{x',y'} I'(x+x',y+y')^2}} \tag{2}$$

where I denotes the image, T – template, $x' = 0...w - 1$, $y' = 0..h - 1$, w is width and h is the height of the template patch used.

The template matching was performed twice on the same image frame I, firstly for nose model as template T secondly with chin model as a template T. The left part of

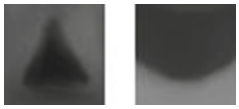

Fig. 4. Nose tip and chin models

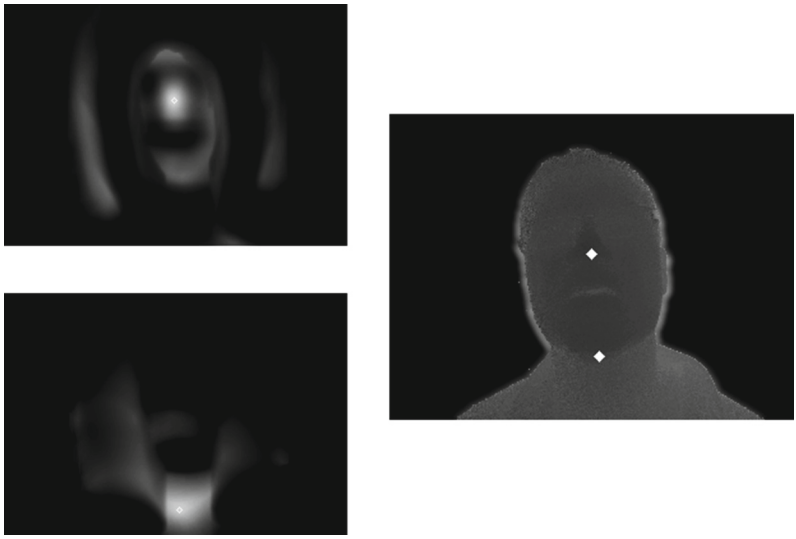

Fig. 5. Nose tip and chin detection results

Fig. 5 depicts the greyscale map showing where the most probable areas of the searched templates are present in the frame. The higher the intensity of the probability map, the larger the value R from Eq. 2 is. The right part of the Fig. 5 presents the result of template matching process with keypoints localized on nose tip and chin. Keypoints are detected in every frame independently, thus the smoothing of the rapid changes of their locations is needed and the Kalman filter [16, 17] was adopted for that purpose.

3.4 Line Fitting and Face Profile View Retrieving

Given the nose tip and chin coordinates, the equation of a line passing through those points can be found. The calculation as in Eq. 3 produces the bisection line of the user's face (which is depicted in Fig. 6):

$$y = \frac{y_2 - y_1}{x_2 - x_1}x + \left(y_2 - \frac{y_2 - y_1}{x_2 - x_1}x_2\right), \tag{3}$$

where x_1, y_1 and x_2, y_2 denote the coordinates of detected chin point and nose tip, respectively. Equation 2 and the locations of the keypoints enable the calculation of the line leading from the bottom to the top of image frame, through the keypoints. Thus, the whole face of the subject is taken into the profile retrieving process.

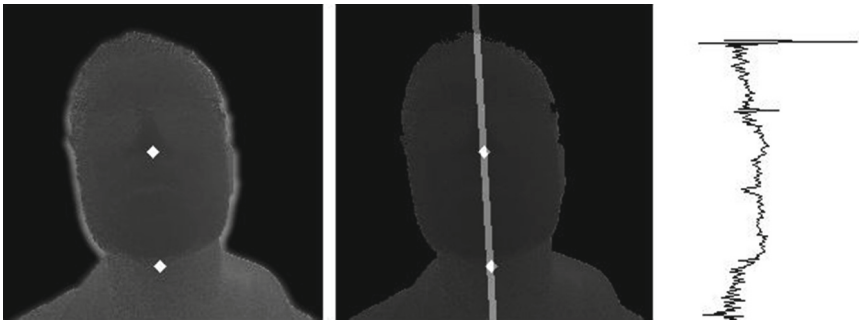

Fig. 6. Result of nose-chin line fitting and profile retrieving from depth data

Having the trajectory of this line and the depth data the profile view of the person can be reconstructed. The width of this line was set to 4 pixels so the final contour will be the mean value of 4 pixel columns from the center of user's face.

The profile view shown in Fig. 6 is a result of retrieving the profile view from the raw input of the TOF camera. This signal is corrupted by the noise, therefore it must undergo a further filtration and modification process in order to extract the characteristic features of considered face. This process is depicted in Fig. 7.

Fig. 7. Filtration of the profile view

The contour signal presented in Fig. 7 left is retrieved from raw depth data the TOF camera provides. The center contour in Fig. 7 is achieved from the depth after median filtration. The right contour is retrieved from median and bilateral filtered. Both median and bilateral filtration are described in a more detailed way in Sect. 3.2 of this paper. The contour shown on the right of Fig. 7, besides of being median and bilateral filtered in 2D domain, provides a subject of 1D median filtering with window size of 5 elements, chosen experimentally. Additionally, the following transformation [1] (Eq. 4) was used in order to emphasize the minima and maxima in the shape of face contour:

$$T = \sqrt{x' + y'} \tag{4}$$

where $x' = x/\max(x)$ and $y' = y/\max(y)$.

3.5 Landmarks Tracking in Profile View

Having the profile view of the subject's face extracted, it is possible to track the facial landmarks by analyzing its minima and maxima. Positions of 9 points in total are tracked i.e. eyebrow, eye, nosetip, nostril, upper and lower lip, corner of mouth, chin depression and chin. For landmarks searching the 1 dimension non maximum suppression method (1D-NMS) is employed. Details concerning the algorithm can be found in the work of Nuebeck and Gool [18]. This method is chosen in order to limit the multiple local minima and maxima which may occur in the close neighborhood while subject utters or changes his or her facial expression. The window size in which two minima or maxima cannot occur is set to 3 pixels. It is the smallest window possible in the 1D-NMS method, nonetheless, in our approach it allows to get stable and accurate landmark positions. The results of landmark searching and tracking are presented in Sect. 4.

4 Results

The image processing methods presented in this paper result in retrieving the filtered signal of subjects' faces contour. In Fig. 8 results of profile view retrieving method application for 4 persons is presented. It is clearly visible that every individual has its own characteristic contour of the face.

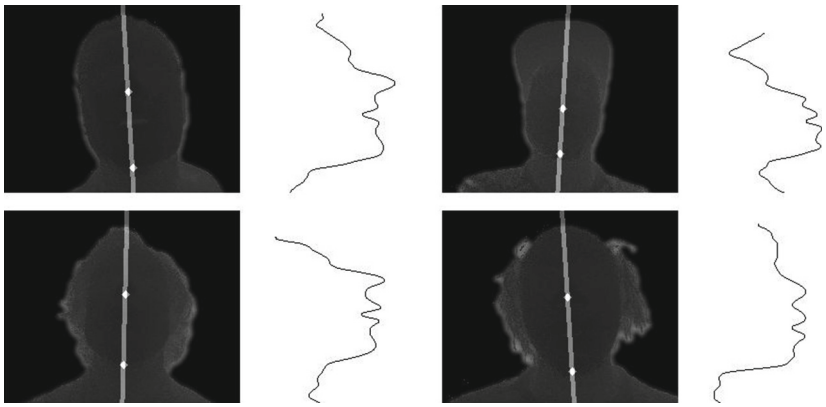

Fig. 8. Four different subjects with corresponding profile views

In Fig. 9 the result of facial landmarks tracking is presented. Thanks to properly filtered shape of the profile view it is possible to search for its minima and maxima positions which correspond with facial landmarks. It is visible that in the three image frames extracted from subject's utterance all 9 landmarks are tracked accurately during the speech. Knowing the landmark positions further parameters may extracted such as geometrical relationships (i.e. static parameters), individual and relative points accelerations (i.e. dynamic parameters), etc.

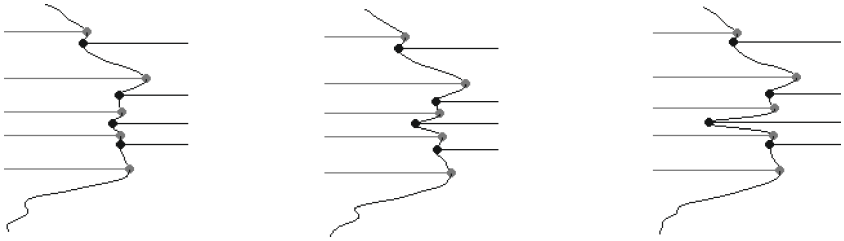

Fig. 9. Profile view facial landmarks tracking during speech

5 Conclusions

The method for profile view of the user's face using the depth data from Time of Flight camera was presented. The method results in filtered contour of the face seen from the profile. The contour signal is algorithmically processed in order to emphasize its distinctive features. The emphasis produces a signal that contains the discriminative features of individual's face. Such a signal may serve as an input to many useful applications, as: Lip-reading in AVSR systems, biometric systems for face recognition, automatic speech therapy systems or facial action recognition systems, to name a few examples.

Additionally, the method of the automatic tracking of facial landmarks locations was presented. Landmarks include 9 characteristic points in total, such as: eyebrows, nose tip, lips, chin depression, chin, etc. This functionality could result in programming API for the future researchers who could use it in the application of their choice which would employ landmarks' coordinates as an input.

Owing to the keypoints tracking of the nose tip and chin and Kalman filtration the proposed method can adapt to subject's movement in front of the device in use. In comparison with other methods that use 3D cameras, our method, thanks to chin and nose tip tracking does not require to have only the subject's face present in the scene as well as the face does not need to be the closest object to the camera. Proposed method is not resource intense and is able to work on-line, with the framerate of 25 fps (on Intel i7 2nd gen processor).

Acknowledgments. This work was supported by the grant No. PBS3/B3/0/2014 Project ID 246459 entitled "Multimodal biometric system for bank client identity verification" co-financed by the Polish National Centre for Research and Development.

References

1. Kumar, K., Chen, T., Stern, R.M.: Profile view lip reading. In: ICASSP (2007)
2. Lucey, P., Potamianos, G.: Lipreading using profile versus frontal views. In: 2006 IEEE 8th Workshop on Multimedia Signal Processing, pp. 24–28 (2006)
3. Pass, A., Zhang, J., Stewart, D.: An investigation into features for multi-view lipreading. In: 17th IEEE International Conference on Image Processing (ICIP 2010), pp. 2417–2420 (2010)

4. Navarathna, R., Dean, D., Sridharan, S., Fookes, C., Lucey, P.: Visual voice activity detection using frontal versus profile views. In: 2011 International Conference on Digital Image Computing Techniques and Applications (DICTA), pp. 134–139 (2011)
5. Dalka, P., Bratoszewski, P., Czyzewski, A.: Visual lip contour detection for the purpose of speech recognition. In: ICSES, Poland (2014)
6. Kunka, B., Kupryjanow, A., Dalka, P., Szczodrak, M., Szykulski, M., Czyzewski, A.: Multimodal english corpus for automatic speech recognition. In: Signal Processing Algorithms, Architectures, Arrangements and Application, Poland (2013)
7. Czyżewski, A., Kostek, B., Ciszewski, T., Majewicz, D.: Language material for english audiovisual speech recognition system development. J. Acoust. Soc. Am. **134**(5), 4069 (2013). (abstr.) plus Proceedings of Meetings on Acoustics, No. 1, vol. 20, pp. 1 – 7, San Francisco, USA, 2.12.2013 – 6.12.2013
8. Zhou, X., Bhanu, B.: Human recognition based on face profiles in video. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2005)
9. Pantic, M., Patras, I.: Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. IEEE Trans. Syst. Man Cyber. 2(2) (2006)
10. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: IEEE International Conference on Computer Vision, Bombay, India (1998)
11. Nagao, M., Matsuyama, T.: Edge preserving smoothing. Comput. Graphics Image Proc. **9**, 394–407 (1979)
12. Yang, N., Kim, Y., Park, R.: Depth hole filling using the depth distribution of neighboring regions of depth holes in the kinect sensor. In: ICSPCC, Honk Kong, pp. 658–661 (2012)
13. Kim, J., Piao, N., Kim, H., Park, R.: Depth hole filling for 3-d reconstruction using color and depth images. In: ISCE, South Korea (2014)
14. Barash, D.: Bilateral filtering and anisotropic diffusion: towards a unified viewpoint. In: Kerckhove, M. (ed.) Scale-Space 2001. LNCS, vol. 2106, pp. 273–280. Springer, Heidelberg (2001)
15. Briechle, K., Hanebeck, U.D.: Template matching using fast normalized cross correlation. In: Proceeding of the SPIE, Optical Pattern Recognition XII, vol. 4387(95) (2001)
16. Kalman, R.E.: A new approach to linear filtering and prediction problems. ASME J Basic Eng. **82**, 35–45 (1960)
17. Szwoch, G., Dalka, P., Czyżewski, A.: Resolving conflicts in object tracking for automatic detection of events in video. Elektronika: konstrukcje, Technologie, zastosowania **52**(1), 52–54 (2011). ISSN 0033-2089
18. Neubeck, A., Van Gool, L.: Efficient non-maximum suppression. In: 18th International Conference on Pattern Recognition (ICPR 2006), vol. 3, pp. 850–855 (2006)