Depth Camera to Improve Segmenting People in Indoor Environments
Real Time RGB-Depth Video Segmentation

Arnaud Boucher[1], Oliver Martinot[2] and Nicole Vincent[3]
[1]*University of Burgundy, Le2i – Route des plaines de l'Yonne – 89000 Auxerre, France*
[2]*Bell Labs, Alcatel Lucent – 91620 Nozay, France*
[3]*Paris Descartes University, LIPADE (SIP) - 45, rue des Saints Peres, 75006 Paris, France*

Keywords: Video Segmentation, Depth+RGB Data, Confidence Estimation.

Abstract: The paper addresses the problem of people extraction in a closed context in a video sequence including colour and depth information. The study is based on low cost depth captor included in products such as Kinect or Asus devices that contain a couple of cameras, colour and depth cameras. Depth cameras lack precision especially where a discontinuity in depth occur and some times fail to give an answer. Colour information may be ambiguous to discriminate between background and foreground. This made us use first depth information to achieve a coarse segmentation that is improved with colour information. Furthermore, color information is only used when a classification in two classes of fore/background pixels is clear enough. The developed method provides a reliable and robust segmentation and a natural visual rendering, while maintaining a real time processing.

1 INTRODUCTION

Segmentation in videos is most often a difficult task and may have different purposes. Some applications answer to editing purposes, such as the matting problem. Most of the solutions still rely on an interaction with the user (Wang, 2007) (Levin, 2008). Many applications are devoted to people tracking, either still or moving people. Tracking according to movement makes the problem easier as some information is added to the only 2D still colour image, based on the comparison of several images. Difficulty is also depending on the camera status, fixed or moving. In any cases, any additional information enables to improve segmentation results. Another way to add information to the colour image is to use several sensors. This approach is more and more frequent as hardware cost is decreasing. The sensors may be based on the same principle, this is the case with stereovision (Jourdheuil, 2012), then depth information can be retrieved from a mathematical model of the coupled sensors, these sensors are now a days available in the field of video and give very good depth precision. The sensors may be based on different spectral information giving complementary information, in some contexts, IR or UV spectrum may be more adapted than visible light. They are used for instance in biometrics applications (Lefevre, 2013) or since a long time in remote sensing domain (Tucker, 1979). In the robotic field, Time of Flight (ToF) cameras are used that give information on the distance of object with respect to the axis of the camera (Wang, 2010). The depth knowledge is integrated to build representation of the objects for robotics tasks (Richtsfield, 2012). Recently, cheap depth camera are coupled with RGB web cam, the quality of these camera is not very high but it is informative enough to stimulate research and develop applications based on this kind of product, the most popular ones being the Kinect camera and the Asus product. Our work is positioned in this trend. The aim is to extract in real time the people present in the scene. The applications are numerous. We mention here as examples, modification of the video background, construction of a two layer video, the person in one layer and the background in the other, augmented reality (Maimone, 2012) (Wu, 2008). We can imagine a speaker in front of his computer where the presentation is displayed and the conference being

Boucher A., Martinot O. and Vincent N..
Depth Camera to Improve Segmenting People in Indoor Environments - Real Time RGB-Depth Video Segmentation.
DOI: 10.5220/0005269700550062
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 55-62
ISBN: 978-989-758-091-8

transmitted with the speaker in foreground and the slide as background. The speaker can show elements on the slide thanks to the feedback on the computer screen. This needs a real time analysis in order to give the speaker the ability to interact with the slides in a virtual way as illustrated in figure 1.

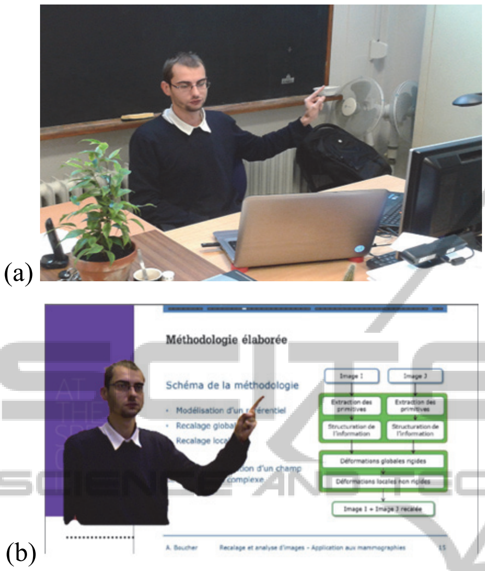


Figure 1: (a) real context, the speaker is looking at the computer screen (b) transferred image where the speaker is incrusted in the presentation in real time.

In section 2, methods applied in foreground / background segmentation and based on the RGB+Depth information are investigated. Next in section 3, the method we developed is analysed and the experiments and results are presented in section 4.

2 RELATED WORKS

In order to use the different types of information, a strategy is needed and makes possible to group the methods that have been proposed. Depth and colour have not the same nature. The difference is twofold, of course the physical nature is not the same but also the resolutions of images are most often different. The combination of the two can be done at different levels, either at raw data level or at feature extraction level or at decision level according to the methods. Using the available data, foreground / background segmentation has to be achieved in a real time process in order to enable interaction with the video content. First we investigate the approach we are interested in before we go more in detail in the segmentation process.

2.1 Characterization of the Approach

Here we consider the global view of the process. Of course, it depends on the acquisition device. Two images, a couple of visual images or a couple of one visual image and one depth image, constitute the material. Then a registration step is needed. It may be performed at hardware level where the two cameras are coupled. In stereovision, from the quality of the registration, is depending the quality of the depth map. With other acquisition devices the quality of registration may vary. Then a registration step has to be integrated in the process (Alempijevic, 2007) (Abramov, 2012).

Then, we distinguish between three ways to process the images.

- The two resulting images may be processed independently, usually leading to the definition of two masks that are combined. The decision may also integrate the past results or data from the previous frames. The final result is obtained after a smoothing as post processing (Camplani, 2014) (Gallego, 2014).
- Another approach is to aggregate the different data in a single data vector. This assumes the quality of the two sensors quality is equivalent otherwise this can give biased results (Stückler, 2010).
- A third approach is to process the two images (RGB and D images) in a sequential way. The first type of data gives a mask that can be improved using the additional information and a post processing.

Taking into account the specificities of the used sensors, we have chosen this last strategy, using depth data to define a coarse mask, improved by the colour information.

2.2 Foreground/background Segmentation

In the literature a wide variety of methods exists to detect an object in RGB and / or Depth images. We here distinguish between methods based on learning and more innovative approaches; the goal is to position and justify our methodological choices.

The learning methods aim to model the characteristics of both background and foreground. The favorite approach is the use of Gaussian mixtures or assumption of other distributions as uniformity in parametric models (Camplani, 2014) (Gallego, 2014) (Schiller, 2011) (Stormer, 2010). Other approaches use non-parametric methods, for example using kernel approaches to model some characteristic distributions (Elgammal, 2002) or a color model based on codebooks as in (Fernandez-Sanchez, 2013). In other cases only raw characteristic samples are stored in order to be used during the decision step (Stückler, 2010). Then, segmentation decision can be handled by a simple thresholding step. This is only efficient when the sensor can efficiently enough discriminate between foreground and background. This is the case with Time on Fly (ToF) sensors (Crabb, 2008) (Wu, 2008) (Frick, 2011). As this type of sensor is very accurate, the processing is simplified. For the sensors with lesser precision we can mention the use of classical clustering methods, such as neural networks (Maddalena, 2008), k nearest neighbour approaches when the learning step is limited to stored samples (KNN) (Barnich, 2011), region growing (Xia, 2011), mean-shift (Bleiweiss, 2009) or random forests (Stückler, 2010). All these well-known clustering methods require training and their efficiency is depending on the type of environment in which the object or people is evolving. Each of the methods is therefore tuned for specific contexts and most only need an adaptive learning.

Other methods attempt to obtain relevant segmentation without a learning step, but based on best boundary between two media. Graphs-cuts are used for instance in a US patent (Do, 2014). From annotations characterizing the inside and outside of the object, these two areas are defined by the minimum cut of the weighted graphs. The orientated graph is characterized by colour and depth differences between neighbouring pixels. This method is efficient when the limits between fore and background are neat and not too complex. One of the difficulties of this method is to fixe the size of the neighbouring zones. Other studies are relying on mean-shift approach that need a lower number of parameters to be tuned.

Depending on the sensors used and the method chosen, the RGB / Depth data can be processed separately, and thus lead to the production of two masks that need to be combined by logical operators. One mask can be considered as master mask that will be improved through the second type data. Indeed we have not discussed the complexity process but this has to be taken into account when real time is needed.

3 OUR METHOD

In our study, data is extracted from low cost devices such as typical camera sensor Kinect or ASUS Xtion and so limited to indoor environments. The segmentation of people must be reliable, efficient, and able to be processed at the rate of 25 frames per second. In our choices, we have taken into account the qualities and defect of the sensor as well as the time limit of 40ms to process a frame. To obtain good results all available information present some interest either for better accuracy or to decrease processing time. As it happens, typical camera sensor such as Kinect or ASUS Xtion, are sold with some more sophisticated capabilities than the output to raw colour or depth signal. In fact the device includes some middleware libraries that are used in the play applications and that are available to the developers. Several information are available such as a pre-segmentation of the individual person present in the scenery, the segmentation of the person in a certain number of significant zones and a skeleton associated with the person. As our aim is not to build an avatar of the person or to analyse his/her gest, but only to reproduce them with high quality of the contours, we have chosen to use only the person extraction module. We have experimented the detection of the persons is robust but the segmentation at the contour level is not very precise. The aim of the work is to improve this segmentation integrating both colour and depth information. The most difficult parts are segmentation of the hairs and precision in segmenting the fingers within the hand when they are not touching.

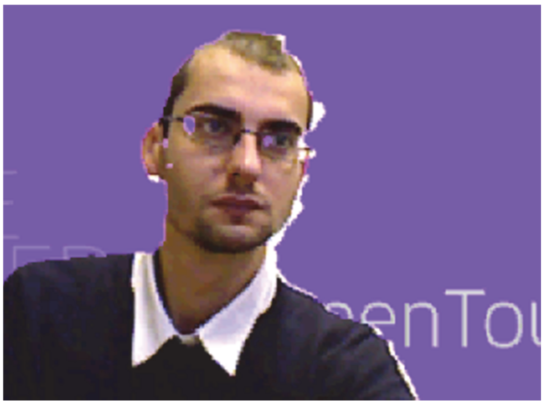


Figure 2 : raw person extraction given by ASUS Xtion software.

3.1 General Organization

Then, in order to spend more time on additional processes, we have chosen to replace the original depth map provided by the camera by the binarised version that is available and is supposed to provide the different persons that are present in the video frame. This is an approximation of the 3D position of the surfaces and includes some parts where the depth has not been measured and where there is no results. An example of such a person extraction is given on figure 2. We also think, the only use of colour information gives results the quality of which is much depending on the image content, on the relative nature of the background and foreground colours. Then, according to the data types, it is better not to process independently the two information sources. Finally, to use the forces of each of the information sources and to prevent their weaknesses we decided to privilege binarised depth map. In fact the resolution of the depth map is lower than the resolution of the colour image. This map can be considered as an initialization of the process that will be improved by incorporating colour data but this is not done uniformly all over the frame but only when colour is pertinent enough for segmenting fore and background. A classification process is here involved.

Possible post processing may allow visual improvement of the final display in order to decrease the imperfection appearances without any correction of the errors. Visual appearance is the most important concern of the study. Next we detail the different steps of the processing (see Figure 3).

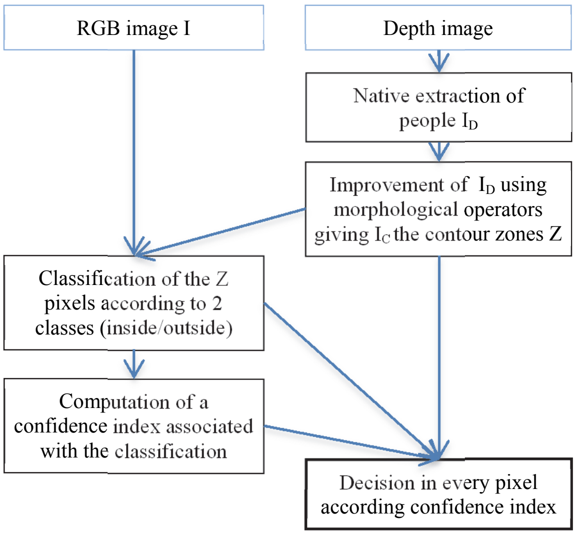

RGB image I
Depth image
Native extraction of people I_D
Improvement of I_D using morphological operators giving I_C the contour zones Z
Classification of the Z pixels according to 2 classes (inside/outside)
Computation of a confidence index associated with the classification
Decision in every pixel according confidence index

Figure 3: General organization.

3.2 Depth Image Processing

The output of ASUS camera system using only depth data provides a coarse segmentation of person(s) detected in the scene. We have chosen to use such information on the fly rather than to process the raw data. This rough segmentation gives a good detection of persons and a rather good estimate of the people contour but with many defaults such as flickers between frames or missing parts of the person. This differentiates our work from (Camplani, 2014) as we have a region approach rather than a contour based approach. The mask associated with presence of a person gives an irregular vision of the person and has to be smoothed using mathematical morphology operators (Serra, 1982; Soille, 1999). In our case we operate a closure with a circular structuring element (radius of 3 pixels) followed by a Gaussian blur to smooth the contour. The visual result is greatly improved, but colour information is not yet used. The result of this first step is a zone supposed to be associated with a person. From this binary image, the contours can be extracted. Next step is to take colour into account to improve the visual rendering.

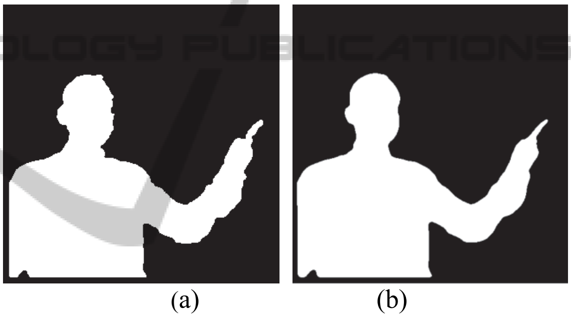
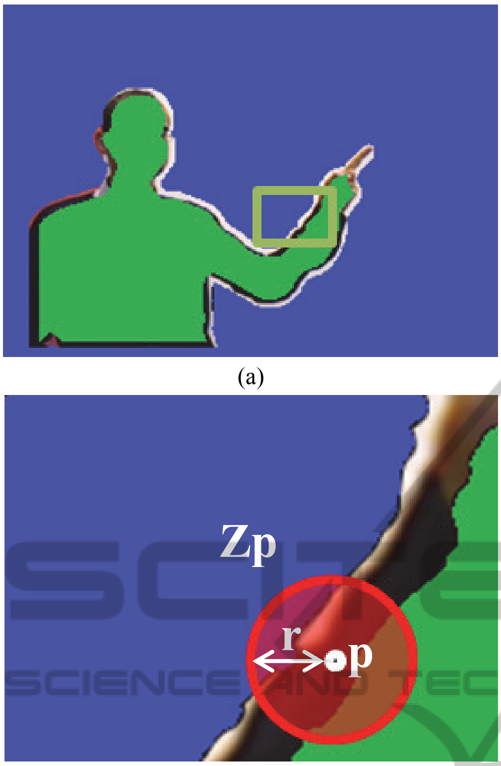


Figure 4: in (a) segmentation issued from the acquisition system (hard and soft) and in (b) improvement after morphological operations.

3.3 Colour Refinement Principle

The refinement is needed only when the contour of the improved depth map are not in coherence with the actual contour of the person to be extracted. This step is crucial for result visual quality. Based on the smoothed depth mask, a study in the neighbour of each contour point is needed; this is to estimate the confidence that can be associated with the depth contour. First we build a trimap like map. The set of contour points is dilated and defines the uncertainty zone. By definition, a priori, the pixels on either sides of the zone are labelled in a confident way as background or speaker (see Figure 5). They can be used to perform a first supervised classification.

Here, the aim is to determine whether the pixel of fore / background can be discriminated and in the positive case to determine the class of each pixel in the uncertainty zone. A global approach, considering



(a)

(b)

Figure 5: In (a) global image with background in dark grey, foreground in light grey and in between an uncertainty zone, in (b) a zoom on the enclosed area in (a), showing the background and the foreground zones specially in the neighbourhood of the white pixel.

the whole image would be defeated when the background or the foreground are not uniform. Then we adopt a local approach in an adaptive way. A neighbouring zone Z_P is associated with each pixel P in such a way that Z_P contains pixels from the background and pixels in the foreground a priori defining two areas B_P and F_P. The colours respectively of the background and foreground can be locally learned. Here they are simply modelled by the mean colours m_{BP} and m_{FP} respectively in B_P and F_P. Thus, pixel P is classified according to its colour C_P and to the mean colour nearest to it (see Figure 5b).

Two distances are compared:

$$d_F(P) = | C_P - m_{FP} | \quad (1)$$

$$d_B(P) = | C_P - m_{BP} | \quad (2)$$

Without any rejection rule in the pixel classification, the result is obtained in a sequential process, first depth processing followed by colour processing. This is not satisfying as all the decisions are based on both types of data in an equal way. The results should be improved if we could benefit only from the best aspects of the two data types. This motivates a new step we present in next session. In fact depth precision is not so high near the edge of the person as there is a rapid change of depth at these points.

3.4 Fusion of Depth and Colour Approaches

As stated in the introduction, as far as our problem is concerned, neither colour nor depth is relevant at all points. Our aim is to combine the information in an intelligent way, locally deciding which information has a better quality and deciding to keep as a final conclusion the result obtained in either the first or the second step of our process. The use of depth initiates the process of a colour-based refinement. Indeed, as shown in figure 6, some edge points are located in areas where speaker and background colours are very similar. In figure 6, the more the colour of the neighbouring zones are red, the more it is difficult to distinguish between fore and background colours. In such cases, colour does not give a discriminant enough information and the confidence in the colour results is low. Thus, colour approach is used only in regions where the colour-based classification is done with high confidence. This confidence is measured at pixels in the uncertainty zone by the difference between the $d_F(P)$ and $d_B(P)$. The colour label is assigned only if this difference is high enough.

$$| d_F(P) - d_B(P) | > s \quad (3)$$

with s a confidence margin. Otherwise the depth image process gives the final conclusion.

The end result is smoothed by a very slight Gaussian blur aimed to eliminate any noise that may prevent misclassification. In next section some analysis of the results is presented.

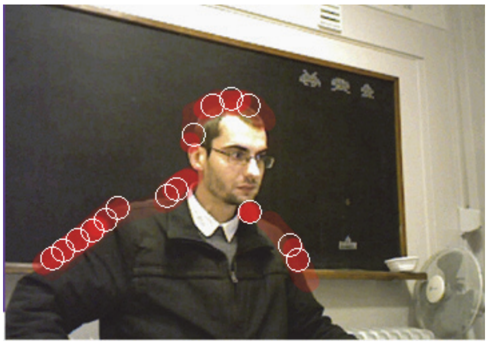


Figure 6: Circles indicate how foreground and background are similar, the more intense red the zone is, the lower the confidence is.

4 EXPERIMENTS AND RESULTS

First of all evaluation protocol has to be set. As precision in the person location is our aim, we have chosen evaluation at pixel level. The measure has to be based on the number of false positive pixels (FP) and of false negative ones (FN). The quality of the results depends on these pixels.

A fair evaluation is very difficult because the different videos do not present the same difficulty level depending on their content, specially the number of people and objects present in the scene, the proximity of these objects, the location of the people in the scene, the similarity between colours in the people and in the background, the rapidity of the people movements, the quality of the data. Then a manually annotated benchmark is needed. We have used many videos with different levels of difficulty. The difference can be seen in Figure 7. The same person makes various movements in two different environments, one quite easy and the other more difficult because of the similarity of the colours of the blackboard and the person coat. A manual segmentation of the person on 5 frames of each video chosen every 100 frames makes the ground truth we use for the quantitative evaluation of the method.

To illustrate the different roles of depth and colour knowledge along the process, we show in figure 8 an image extracted from one of the video both the RGB image (8a) and the depth image (8b) as well as the ground truth (8f). Then we have illustrated the difference between the ground truth and the segmentation results obtained from the depth map smoothed by morphological operations (8c), the improvement using colour (8d) in a uniform way along the contour, and the final result taking into account the context (8e), that is choosing according to the confidence associated with the local contrast between fore and background. The quality is linked to ratio of grey pixels. The false positive and the false negative pixels are respectively indicated in white and black. The grey level indicates the segmentation is coherent with the ground truth.

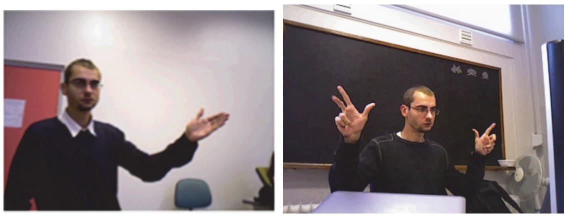
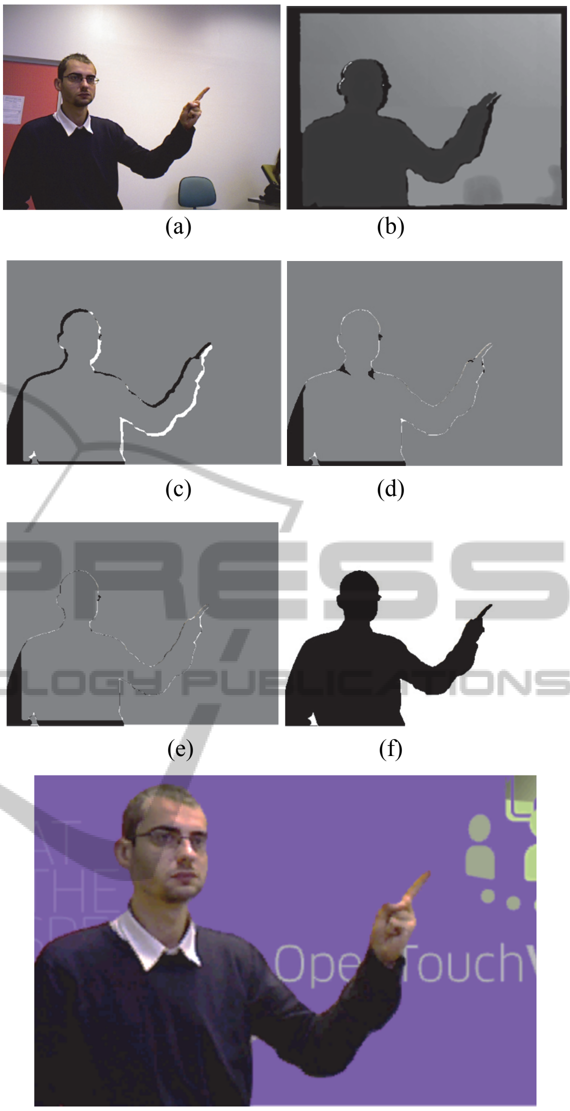

(a) (b)

Figure 7: two videos of different difficulties, in (a) high contrast between the person and the background and in (b) a more difficult on with similar colour on the person and the background.


(a) (b)

(c) (d)

(e) (f)

Figure 8: In (a) and (b) the initial RGB and Depth images, in (c) results using only depth information, in (d) the improvement based on colour, in (e) the combination of the two, in (f) the ground truth segmentation, (g) natural segmentation based on (e).

On this example, we can notice the good improvement at the finger level. Also on the depth image we can see the bad segmentation at the hair level. The top part of the door is mistaken with the person hair. At the bottom, the arm is not differentiated from the body.

The quantitative evaluation is based on the 5 images extracted in the videos. Recall, precision of the pixels of the person are classical indexes to assess the method. In figure 9 and figure 10, we are

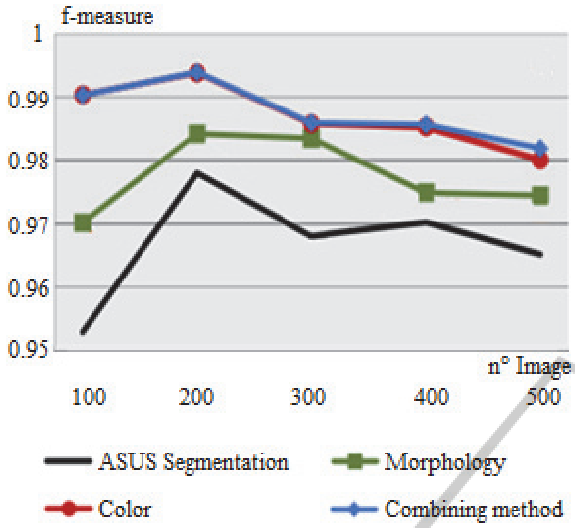


Figure 9: F-measure of proposed segmentation on the different frames with the "easy" video where the contrast is high between the person and the background, illustrated in figure 7a.

presenting the F-measure associated with the original result provided by the camera itself, the smoothing morphology, the improvement given by colour, the adaptive combining method.

$$F - measure = \frac{2\,R\,F}{R+P} \qquad (4)$$

with $R = \frac{TP}{TP+FN}$ and $P = \frac{TP}{TP+FP}$

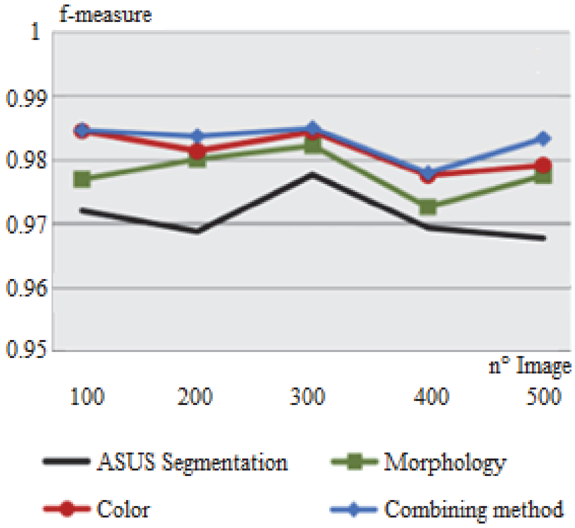


Figure 10: F-measure of proposed segmentation on the different frames with the "difficult" video where the person and the background have nearly the same colour, the video is illustrated in figure 7b.

The results show that simple morphology improves result. Nevertheless, colour enables to have good results in "easy" cases but fails when the video is more "difficult". Finally combining the two sources of information in an adaptive way thanks to the notion of uncertainty, gives the better results. The results are obtained in real-time on an Intel Pentium i7.

5 CONCLUSIONS

In this paper, relying on the initial results provided by the commercial new sensors based on RGB-D data, we introduced a real time and robust method using information produced by a new type of sensor in an adaptive way, in an intelligent way. The results applied to the segmentation of a speaker in a video-conference context are convincing both on the visual aspect and on the computation time. The growing power of computers will enable to improve the colour processing that is to say the classification of the pixels, here we model the colours locally by a mean value. Both the F-measure computed and the user feedbacks show that the visual result is quite satisfactory. Further evaluation has to be performed and the method could be tested with other sensors. Nevertheless, visual enhancements may be introduced to make the display of the results in an even more agreeable visual perception. A smoothing process can achieve this. The flicker aspect can be eliminated making the process depend on the movement evaluation of the person. When there is no movement, the previous result can be considered as well calculated, when movement is occurring a new calculus is more appropriated. A combination depending on the movement speed is a good solution.

REFERENCES

Abramov A., Pauwels K., Papon J., Worgotter F., Babette Dellen B., 2012. Depth-supported real-time video segmentation with the kinect, *in Workshop on the Applications of Computer Vision.*

Alempijevic A., Kodagoda S., Dissanayake G.,2007. Sensor Registration for Robotic Applications, in proceedings of 6^{th} *International Conference on Field and Service Robotics*, Chamonix, France.

Barnich O., Van Droogenbroeck M., 2011. ViBe: a universal background subtraction algorithm for video sequences, In *IEEE Transactions on Image Processing*, vol 20(6), pp. 1709–1724.

Bleiweiss A., Werman M., 2009. Fusing time-of-flight depth and color for real-time segmentation and tracking, In *Conference on Dynamic 3D Imaging*, pp. 58–69.

Camplani M., Salgado L., 2014. Background foreground segmentation with RGB-D Kinect data: An efficient combination of classifiers, *Journal of Visual Communication and Image Representation*, vol 25(1), pp. 122-136.

Crabb C., Tracey R., Puranik A., Davis J., 2008. Real-time foreground segmentation via range and color imaging", In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–5.

Do M., Lin D., Meyer G., Nguyen Q., Patel S., 2014. Systems and methods for accurate user foreground video extraction, *U.S. Patent Application* 13/083,470.

Elgammal A., Duraiswami R., Harwood D., Davis L., 2002. Background and foreground modeling using non-parametric kernel density estimation for visual surveillance, In *Proceedings of the IEEE*, vol 90, pp.1151–1163.

Fernandez-Sanchez E., Diaz J., Ros E., 2013. Background Subtraction Based on Color and Depth Using Active Sensors. Sensors 13 (7), p. 8895-8915.

Frick A., Franke M., Koch R., 2011. Time-consistent foreground segmentation of dynamic content from color and depth video, Pattern Recognition, Elsevier, pp. 296–305.

Gallego J., Pardas M., 2014. Region based foreground segmentation combining color and depth sensors via logarithmic opinion pool decision, *Journal of Visual Communication and Image Representation*, vol 25(1), pp.184-194.

Jourdheuil L., Allezard N., Château T. and Chesnais T., 2012. Heterogeneous adaboost with realtime constraints - application to the detection of pedestrians by stereovision, In *Proceedings VISAPP'12*, pp. 539–546.

Lefevre T., Dorizzi B., Garcia-Salicetti S., Lempérière N., Belardi S., 2013. Effective elliptic fitting for iris normalization. *Computer Vision and Image Understanding* 117(6): 732-745.

Levin, A., Rav-Acha, A., Lischinski, D., 2008. Spectral matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(10), pp.1699–712.

Maddalena I., Petrosino A., 2008. A self-organizing approach to background subtraction for visual surveillance applications, IEEE Transactions on Image Processing 17, pp. 1168–1177.

Maimone A., Bidwell J., Peng K., Fuchs H., 2012. Enhanced personal autostereoscopic telepresence system using commodity depth cameras. *Computers & Graphics* 36 (7) p. 791-807.

Richtsfield A., Morwald T., Prankl J., Balzer J., 2012. Towards scene understanding – object segmentation using using RGBD-images, in Computer Vision Winter Workshop.

Schiller I., Koch R., 2011. Improved video segmentation by adaptive combination of depth keying and mixture-of-Gaussians, *Image Analysis*, pp. 59–68.

Serra J., 1982. Image Analysis and Mathematical Morphology, Academic Press, London.

Soille P., 1989. Morphological Image Analysis : Principles and Applications. Springer-Verlag.

Stormer A., Hofmann M., Rigoll G., 2010. Depth gradient based segmentation of overlapping foreground objects in range images, In *proceedings of IEEE 13th Conference on Information Fusion*, pp.1–4.

Stückler J., Behnke S., 2010. Combining depth and color cues for scale and Viewpoint Invariant object segmentation and recognition using Random Forests, In *proceedings International Conference on Intelligent Robots and Systems* (IROS), pp. 4566-4571.

Tucker C., 1979.Red and photographic infrared linear combinations for monitoring vegetation, Remote Sensing of Environment Volume 8, Issue 2, May 1979, Pages 127–150.

Wang J., Cohen M., 2007. Image and Video Matting : A Survey. *Computer Graphics and Vision*, pp.1–78.

Wang L., Zhang C., Yang R., Zhang C., 2010. Tof cut: towards robust real-time foreground extraction using a time-of-flight camera, Conference 3D PVT.

Wu Q., Boulanger P., Bischof W., 2008. Robust real-time bi-layer video segmentation using infrared video, In *proceedings of Conference on Computer and Robot Vision* (CRV), pp. 87–94.

Xia L., Chen C., Aggarwal J., 2011. Human detection using Depth information by Kinect, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 15-22.