MULTI-MODAL BACKGROUND SUBTRACTION USING GAUSSIAN MIXTURE MODELS

Benjamin Langmann, Seyed E. Ghobadi, Klaus Hartmann, Otmar Loffeld

ZESS - Center for Sensor Systems, University of Siegen, Germany, (langmann, ghobadi, hartmann, loffeld)@zess.uni-siegen.de

Commission III, WG III/5

KEY WORDS: Classification, Detection, Scene, Segmentation, Video

ABSTRACT:

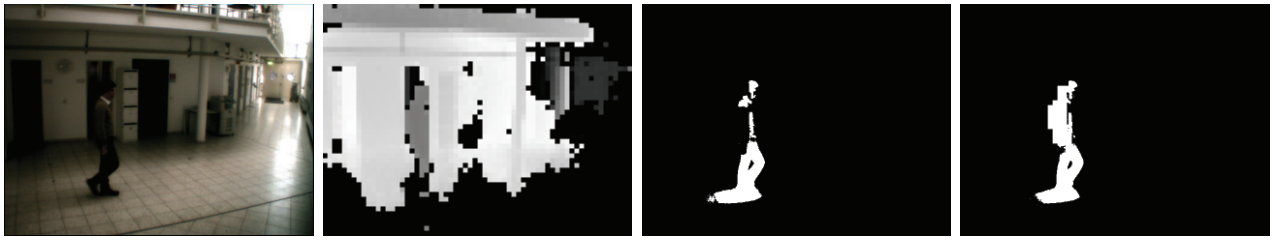
Background subtraction is a common first step in the field of video processing and it is used to reduce the effective image size in subsequent processing steps by segmenting the mostly static background from the moving or changing foreground. In this paper previous approaches towards background modeling are extended to handle videos accompanied by information gained from a novel 2D/3D camera. This camera contains a color and a PMD chip which operates on the Time-of-Flight operating principle. The background is estimated using the widely spread Gaussian mixture model in color as well as in depth and amplitude modulation. A new matching function is presented that allows for better treatment of shadows and noise and reduces block artifacts.
Problems and limitations to overcome the problem of fusing high resolution color information with low resolution depth data are addressed and the approach is tested with different parameters on several scenes and the results are compared to common and widely accepted methods.

1 INTRODUCTION

For a few years cameras utilizing PMD technology to gain depth information through the Time-of-Flight principle have been available. These ToF cameras are often combined with standard color cameras. Recently, 2D/3D cameras which contain a color and a PMD chip in a monocular setup have been developed (MultiCam (Prasad et al., 2006), ZCam), which eliminates the need for a registration of the produced images. These cameras supply besides a color and a depth image also an amplitude modulation image, which indicates how much infrared light was received by the PMD chip. Apart from their advantages of high frame rates and their ability to capture the scene all at once, all PMD based cameras have also the disadvantages of a low resolution, typically 64×48 up to 204×204, high noise and difficulties lighting large scenes. Nevertheless, the 2D/3D cameras supply additional dimensions compared to ordinary video which makes it possible to overcome ambiguities and distortions in standard image processing tasks.
The focus of this paper lies in the common image processing step of background subtraction, modeling or segmentation. The goal is to isolate regions of interest in each image, which are here defined as the the moving foreground opposed to the mostly static background of the video. Methods to that end can be divided coarsely into two classes: pixel- and region-based approaches. The earlier perform the classification of a pixel based only on already known information about that pixel, whereas the later use information from neighboring pixels (often grouped in regions or layers) as well. The most notable of the pixel-based approaches is the method of Gaussian Mixture Models with is widely spread, simple and efficient. The background is hereby modeled by a mixture of Gaussians for each pixel - hence the name. When the observations are accompanied by a depth value and an amplitude modulation for each pixel, the problem of how to combine these different types of dimensions in the classification step has to be addressed.
In this work the background subtraction method based on Gaussian Mixture Models (GMM) is adapted to videos with color, depth and amplitude modulation gained through the Time-Of-Flight principle, which will be referred to as 2D/3D videos (see figure 1 for an example). Here the significantly lower resolutions of the depth and amplitude modulation images have to be accounted for additionally. To that end a measure that links the dimensions in a statistical manner is presented which also results in a lower noise level and a better classification of shadows. Previous methods either use rather simple foreground estimation methods or are designed to operate on full sized depth images which were gained by a stereo setup.

2 RELATED WORK

In (Ghobadi et al., 2008) the foreground of 2D/3D videos is extracted simply by defining a volume of interest and this is used for hand tracking as well as gesture recognition. Harville et al. applied the standard approach of background modeling by Gaussian mixtures, see e.g. (Stauffer and Grimson, 1999), to color and depth videos in (Harville et al., 2001). They are using full sized depth images so that there is no need to handle the different resolutions.
In (Bianchi et al., 2009) a rather simple approach to foreground segmentation for 2D/3D videos that is based on region growing and refrains from modeling the background is evaluated, whereas in (Leens et al., 2009) a simple pixel-based background modeling method called ViBe is used for color and depth dimensions separately and the resulting foreground masks are fused with the help of binary image operations such as erosion and dilation.
A more elaborate method of fusing color and depth is bilateral filtering, which is used e.g. in (Crabb et al., 2008). Here the preliminary foreground is produced by a dividing plane in space and a bilateral filter is applied to gain the final results. The method is demonstrated on depth augmented alpha matting, which is also the focus of the paper (Wang et al., 2007). In (Schuon et al., 2008) the ability of bilateral filtering to deal with geometric objects is demonstrated and in (Chan et al., 2008) a variant designed to handle noise and invalid measurements is presented.
The problem of fusing the depth and color dimensions and handling their different nature is also discussed in the course of depth upscaling. To that end a cost function or volume is defined in (Yang et al., 2007), that describes the cost of in theory all possi-



(a) Input video (b) Depth image (c) GMM on color (d) GMM on color and depth

Figure 1: Results for (a) a challenging input video , (b) a depth image where invalid measurements are black, (c) the foreground mask using GMM only for color and (d) GMM based on color and depth

ble refinements of the depth for a color pixel. Again a bilateral filter is applied to this volume and after sub-pixel refinement a proposed depth is gained. The optimization is performed iteratively to achieve the final depth map. The incorporation of a second view is also discussed. In (Bartczak and Koch, 2009) a similar method using multiply views was presented.
An approach working with one color image and multiple depth images is described in (Rajagopalan et al., 2008). Here the data fusion is formulated in a statistical manner and modeled using Markov Random Fields on which an energy minimization method is applied.
Another advanced method to combine depth and color information was introduced in (Lindner et al., 2008). It is based on edge preserving biquadratic upscaling and performs a special treatment of invalid depth measurements.

3 GAUSSIAN MIXTURE MODELS

All observations at each pixel position $\underline{x} = [x, y]^T$ are modeled with a mixture of Gaussians to estimate the background. The assumption is that an object in the line of sight associated with a certain pixel produces a Gaussian formed observation or several in the case of a periodically changing appearance (e.g., moving leaves, monitor flickering). Each observation is then modeled with one Gaussian whose mean and variance is adapted over time. An observation at time t for a pixel $\underline{x}$ is given by $\underline{s}^t(\underline{x}) = [s_1^t(\underline{x}), s_2^t(\underline{x}), \ldots, s_n^t(\underline{x})]^T$. The probability distribution density of $\underline{s}^t(\underline{x})$ can now be described by

$$f_{\underline{s}^t(\underline{x})}(\underline{\xi}) = \sum_i \omega_i \cdot N_{\underline{\xi}}\left(\underline{\mu}_i, \Sigma_i\right) \tag{1}$$

where

$$N_{\underline{\xi}}\left(\underline{\mu}_i, \Sigma_i\right) = \left[(2\pi)^{\frac{\dim(s)}{2}} \cdot \det(\Sigma_i)^{\frac{1}{2}}\right]^{-1} \cdot \exp\left\{-\frac{1}{2}\left[\underline{\xi} - \underline{\mu}_i\right]^T \cdot \Sigma_i^{-1} \cdot \left[\underline{\xi} - \underline{\mu}_i\right]\right\} \tag{2}$$

is the multivariate Gaussian with mean $\underline{\mu}_i$ and covariance matrix Σ_i. Clearly for the mixing coefficients ω_i we must have: $\sum_i \omega_i = 1$. How many Gaussians should be used to model the observations, how to adapt the Gaussian efficiently over time and which Gaussians should be considered background, are questions that arrive immediately. Most GMM based methods are based on very simple assumptions for efficiency. They used a fixed number of Gaussians per pixel and the minimum number of Gaussians with weights which sum up to a given threshold are treated as background.
The adaptation of the Gaussians over time is a bit more complicated. Instead of using the EM-algorithm or similar methods to determine the parameters of the mixture, the usual online clustering approach is used in this work: When a new observation $\underline{s}(\underline{x})$ arrives it is checked if it is similar to already modeled observations or if it is originating from a new object. It may also just be noise. This is done by evaluating the Mahalanobis distance $\delta(\cdot,\cdot)$ towards the associated Gaussian $N_{\underline{x}}(\underline{\mu}_i, \Sigma_i)$

$$\delta\left(\underline{x}, \underline{\mu}_i\right) = \sqrt{\left(\underline{\mu}_i - \underline{x}\right)^T \Sigma_i^{-1} \left(\underline{\mu}_i - \underline{x}\right)} < T_{near} \tag{3}$$

with T_{near} being a given constant. If similar observations have been recorded, their Gaussian is adapted using the observed data. Otherwise, a new Gaussian is created and added to the mixture. An exact description of a possible implementation can be found in (Stauffer and Grimson, 1999) for normal videos and in (Harville et al., 2001) with additional depth values.
An observation for a pixel is given by $\underline{s}(\underline{x}) = (y, c_b, c_r, z, a)^T$ in this work and contains the color value in YCbCr format, a depth value z and an amplitude modulation value a. The 2D/3D camera produces a full size color image and low resolution depth and amplitude modulation images which are resized to match to color images by the nearest neighbor method. The variances of all Gaussians are limited to be diagonal to simplify computations. When working with ToF data, invalid depth measurements due to low reflectance, have to be handled cautiously. A depth measurement is considered invalid if the corresponding amplitude is lower that a given threshold. In (Harville et al., 2001) an elaborate logical condition is used to classify a new observation. Experiments show that this can be simplified by using the measure

$$\widehat{\delta}\left(\underline{x}, \underline{\mu}_i\right)^2 = \left(\underline{\mu}_i - \underline{x}\right)^T \Sigma_i^{-1} \begin{pmatrix} 1 & & & \\ & \lambda_c & & \\ & & \lambda_c & \\ & & & \lambda_z \\ & & & & \lambda_a \end{pmatrix} \left(\underline{\mu}_i - \underline{x}\right) \tag{4}$$

and checking the condition

$$\widehat{\delta}\left(\underline{x}, \underline{\mu}_i\right)^2 < T_{near}^2 \cdot Tr\begin{pmatrix} 1 & & & \\ & \lambda_c & & \\ & & \lambda_c & \\ & & & \lambda_z \\ & & & & \lambda_a \end{pmatrix} = T_{near}^2 \left(1 + 2\lambda_c + \lambda_z + \lambda_a\right) \tag{5}$$

where $\lambda_z \in \{0, 1\}$ depending on whether current and previous depth measurements are both valid. The mechanism from (Harville et al., 2001) works well to that end. Similarly, $\lambda_c \in \{0, 1\}$ indicates whether the chromaticity channels of the current observation as well as the recorded information provide trustworthy values. This can be estimated simply by checking if both luminance values or their means respectively are above a certain threshold. Finally, $\lambda_a \in \{0, 1\}$ determines if the amplitude modulation should be used for the classification and it is specified a priori.
This matching function pays respect to the fact that observations in the color, depth and amplitude modulation dimensions are in practice not independent. A foreground object has most likely not

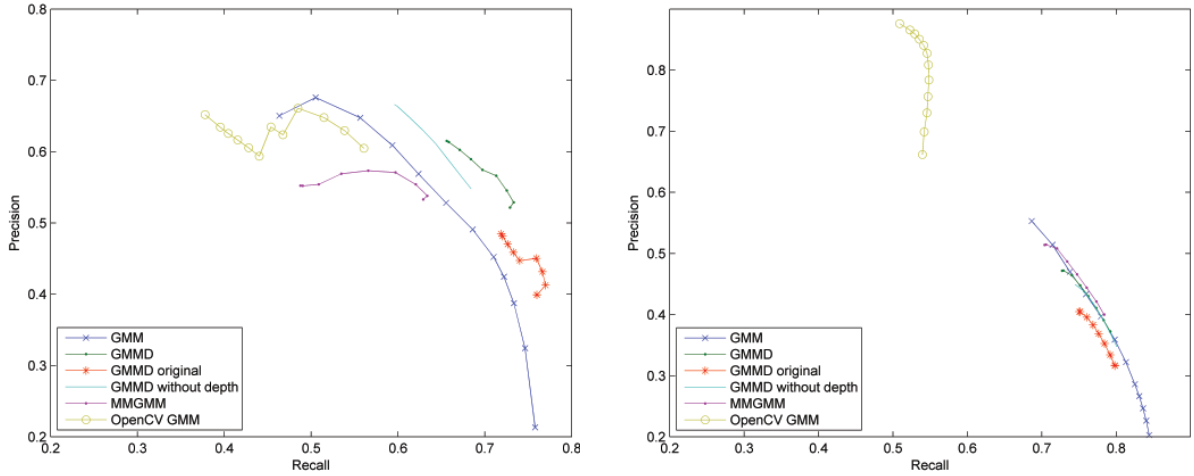


Figure 2: Average recall and precision values for different background subtraction methods using different parameters. Left: values for the 2D/3D video shown in table 1, right: for the video shown in table 2

only a different depth but also at least a slightly different color and infrared reflectance properties. Other reasons are limitations of video and ToF cameras, e.g., the infrared reflectance of an object has an influence on the depth measurement (or its noise level) and low luminance impairs chromaticity measurements. Therefore, a linkage between the dimensions reduces the noise level in the foreground mask, the amount of misclassification due to shadows and block artifacts which occur when only depth measurements are inappropriate.
More elaborate variants such as learning modulation and special treatment of deeper observations when determining what observations are considered background are described in (Harville et al., 2001) but do not seem to be necessary for simple scenarios.

4 EXPERIMENTS

The approach described in this work can be evaluated by examining if it obeys the following principles:

- When the ordinary background subtraction based on color works, the results should not be harmed, e.g., through block artifacts at the border of the foreground mask.
- When the foreground is not classified correctly only with color, this should be compensated by depth.
- The shadow treatment of the color based background subtraction is still far from perfect and should be improved through depth information.

The following methods were compared in the course of this work. 'GMM' is the standard color based GMM approach (Stauffer and Grimson, 1999) and 'Original GMMD' is the original color and depth based method from (Harville et al., 2001). 'GMMD without depth' is the method described in this work without depth measurements (always $\lambda_z = 0$) and with $\lambda_a = 0$, whereas in 'GMMD' λ_z is determined based on the amplitude modulation for each pixel similar as in (Harville et al., 2001) and in 'MMGMM' $\lambda_a = 1$ is set additionally. The values for the OpenCV GMM method are given for reference only, since it contains post-processing steps and is therefore not directly comparable.
In table 1 the results for a 2D/3D video with difficult lighting conditions using these methods are shown. The same parameters were used for all methods: a maximum number of 4 Gaussians per pixel, a learning rate of $\alpha = 0.0005$, an initial $\sigma = 5$ and a threshold $T_{near} = 3.5$. Due to the fact that all methods operate based on the same principle the results should be comparable for a given set of parameters. This was also confirmed by several parameter variations.
The results demonstrate the ability of this method to achieve the mentioned objectives. The misclassification of shadows is reduced and the natural borders of the foreground are harmed less. When the classification based on color fails, these areas are filled at least partly. The compensation is unfortunately often done in a blockwise fashion (see figure 1). This drawback is further discussed the next section.
Image sequences from another experiment are shown in table 2 using the same parameter set. Here the lighting conditions are far better so that the standard GMM algorithm can in theory distinguish between foreground and background. On the other hand shadows and the similarity between foreground (jacket) and background cause large problems in this video. The method proposed in this work does not affect the good classification based on color but allows for better shadow treatment due to the available depth values.
In figure 2 quantitative results for both 2D/3D videos are shown. A ground truth was created per hand for every 5th frame starting with the last empty frame before the person enters the scene and ending with with first empty frame after the person has left the scene. Then the number of true positives tp, false positives fp and false negatives fn was counted in each frame for the different methods using thresholds $T_{near} = 2, 2.5, \ldots, 8$ to calculate the recall $tp/(tp+fn)$ and the precision $tp/(tp+fp)$ values and their average over all frames was plotted. Here all variants of the proposed methods perform superior to the classic approach and to the original GMMD method with the exception of the MMGMM method in the first video which on the other hand achieves the best results for the second video. This behavior is due to the fact that the scene in video 1 is much more difficult to light than the scene from video 2, which results in higher noise levels in the amplitude modulation images in video 1. The very different values for the OpenCV GMM method for the second video are caused by the fact that this method classifies the TV correctly, whereas all other methods fail in that respect. The comparably low recall values, i.e., a largely incomplete true foreground possibly due to the foreground background similarity, for the OpenCV GMM method are worth mentioning.

5 LIMITATIONS

The ordinary color based GMM background subtraction cannot distinguish between foreground and background when the color difference is small due to its pixel based nature. The depth values gained from a ToF camera provide the ability for a correct classification of all image blocks with depth values different from those of the background as long as there are valid depth measurements for the background. As illustrated in figure 3 classification based only on low resolution depth values will result in an unnatural foreground contour due to block artifacts. Practically, this drawback cannot be resolved in typical situations, because in such areas the background usually continues with the same color so that there is no edge that would allow gradient based methods to smooth the contour of the foreground mask correctly. Otherwise, bilateral filtering, see (Tomasi and Manduchi, 1998), which is often used in the context of 2D/3D videos to enhance the resolution in depth, would be able to reconstruct the true contour of the object.
To resolve the general case contour estimation methods that incorporate knowledge of the object given a priori or learned through time are necessary, but it does not seem to be possible to achieve good results in a not strictly defined setting. Only in the opposite case, when inappropriate depth measurements result in a wrong classification, gradient based methods can be applied to smooth the contour.

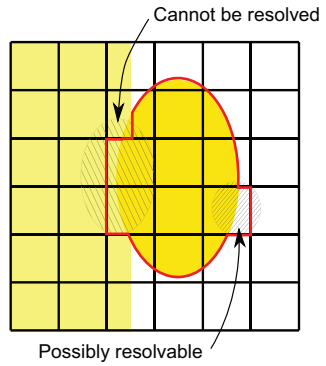


Figure 3: Illustration of a foreground mask. Dark yellow: foreground object, light yellow: background with similar color, red: detected object contour

6 CONCLUSION

In this paper the standard method for background subtraction based on Gaussian Mixture Models is adapted to operate on videos acquired with a 2D/3D camera. The proposed method was compared to standard and previous methods using simple 2D/3D video sequences. Qualitative as well as quantitative results were presented and it was found that the proposed method is able to compensate for misclassification of pixels due to color similarities between foreground objects and the background by utilizing depth and modulation amplitude information without harming the high resolution contour of foreground objects. Furthermore, this method provides a clearly improved treatment of shadows and noise compared to previous methods.
The additional burden compared with standard background subtraction methods based on GMM to process and maintain the depth values is small, i.e., on current PCs real-time processing is easily possible.

ACKNOWLEDGEMENTS

This work was funded by the German Research Foundation (DFG) as part of the research training group GRK 1564 'Imaging New Modalities' and the authors would like to thank Omar E. Löpprich for the help with recording the 2D/3D videos and the valuable discussions.

REFERENCES

Bartczak, B. and Koch, R., 2009. Dense depth maps from low resolution time-of-flight depth and high resolution color views. In: Proc. of ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications, Lecture Notes in Computer Science, Vol. 5876, pp. 228–239.

Bianchi, L., Dondi, P., Gatti, R., Lombardi, L. and Lombardi, P., 2009. Evaluation of a foreground segmentation algorithm for 3d camera sensors. In: ICIAP, Lecture Notes in Computer Science, Vol. 5716, Springer, pp. 797–806.

Chan, D., Buisman, H., Theobalt, C. and Thrun, S., 2008. A noise-aware filter for real-time depth upsampling. In: Proc. of ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications.

Crabb, R., Tracey, C., Puranik, A. and Davis, J., 2008. Real-time foreground segmentation via range and color imaging. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08., pp. 1–5.

Ghobadi, S. E., Loepprich, O. E., Ahmadov, F., Bernshausen, J., Hartmann, K. and Loffeld, O., 2008. Real time hand based robot control using 2d/3d images. In: ISVC '08: Proceedings of the 4th International Symposium on Advances in Visual Computing, Part II, Springer-Verlag, Berlin, Heidelberg, pp. 307–316.

Harville, M., Gordon, G. and Woodfill, J., 2001. Foreground segmentation using adaptive mixture models in color and depth. In: Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video, IEEE Computer Society, Los Alamitos, CA, USA, pp. 3–11.

Leens, J., Piérard, S., Barnich, O., Droogenbroeck, M. V. and Wagner, J.-M., 2009. Combining color, depth, and motion for video segmentation. In: ICVS '09: Proceedings of the 7th International Conference on Computer Vision Systems, Springer-Verlag, pp. 104–113.

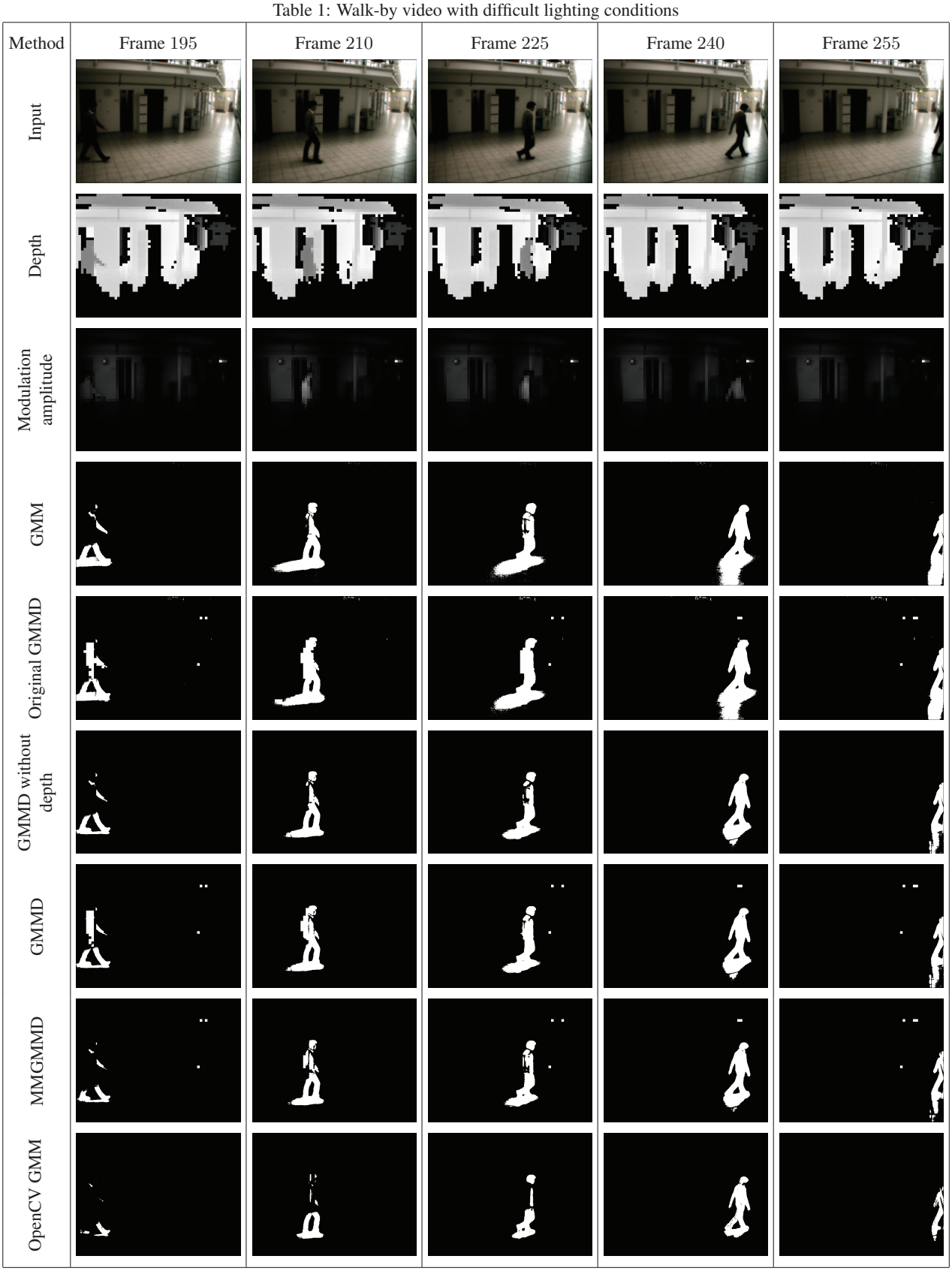
Lindner, M., Lambers, M. and Kolb, A., 2008. Sub-pixel data fusion and edge-enhanced distance refinement for 2d/3d images. Int. J. Intell. Syst. Technol. Appl. 5(3/4), pp. 344–354.

Prasad, T., Hartmann, K., Wolfgang, W., Ghobadi, S. and Sluiter, A., 2006. First steps in enhancing 3d vision technique using 2d/3d sensors. In: 11. Computer Vision Winter Workshop 2006, Czech Society for Cybernetics and Informatics, University of Siegen, pp. 82–86.

Rajagopalan, A. N., Bhavsar, A., Wallhoff, F. and Rigoll, G., 2008. Resolution enhancement of pmd range maps. In: Proceedings of the 30th DAGM symposium on Pattern Recognition, Springer-Verlag, pp. 304–313.

Schuon, S., Theobalt, C., Davis, J. and Thrun, S., 2008. High-quality scanning using time-of-flight depth superresolution. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on, pp. 1–7.

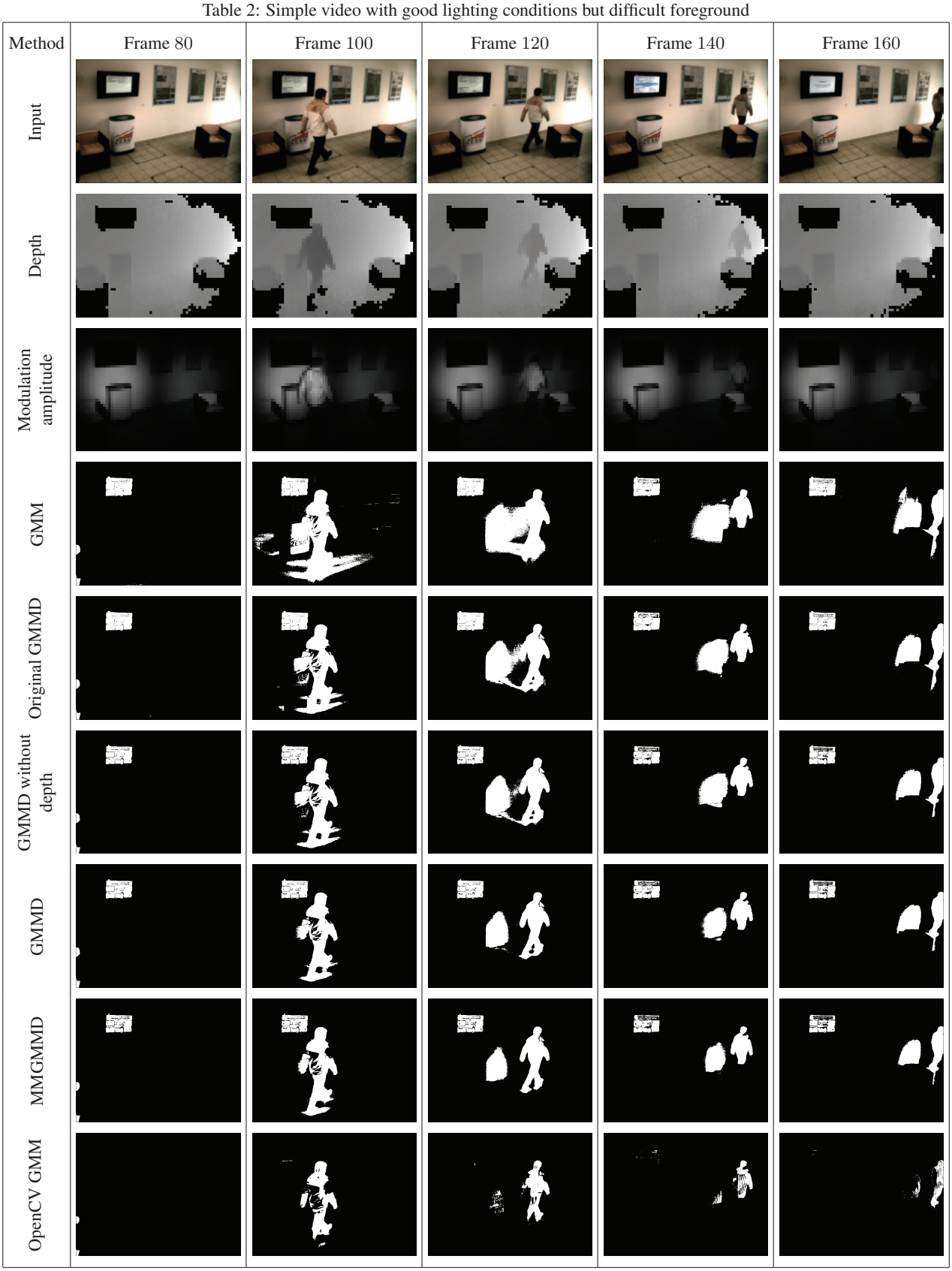
Table 1: Walk-by video with difficult lighting conditions



Stauffer, C. and Grimson, W. E. L., 1999. Adaptive background mixture models for real-time tracking. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2, pp. 246–252.

Tomasi, C. and Manduchi, R., 1998. Bilateral filtering for gray and color images. In: ICCV '98: Proceedings of the Sixth International Conference on Computer Vision, IEEE Computer Society, Washington, DC, USA, pp. 839–846.

Table 2: Simple video with good lighting conditions but difficult foreground



Wang, O., Finger, J., Yang, Q., Davis, J. and Yang, R., 2007. Automatic natural video matting with depth. In: PG '07. 15th Pacific Conference on Computer Graphics and Applications, pp. 469–472.

Yang, Q., Yang, R., Davis, J. and Nister, D., 2007. Spatial-depth super resolution for range images. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pp. 1–8.