FOREGROUND SILHOUETTE EXTRACTION ROBUST TO SUDDEN CHANGES OF BACKGROUND APPEARANCE

Alexandre Alahi, Luigi Bagnato, Damien Matti and Pierre Vandergheynst

Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

Vision-based background subtraction algorithms model the intensity variation across time to classify a pixel as foreground. Unfortunately, such algorithms are sensitive to appearance changes of the background such as sudden changes of illumination or when videos are projected in the background. In this work, we propose an algorithm to extract foreground silhouettes without modeling the intensity variation across time. Using a camera pair, the stereo mismatch is processed to produce a dense disparity based on a Total Variation (TV) framework. Experimental results show that with sudden changes of background appearance, our proposed TV disparity-based extraction outperforms intensity-based algorithms and existing stereo-based approaches based on temporal depth variation and stereo mismatch.

Index Terms— Background subtraction, foreground silhouettes, total variation, stereo camera, disparity map.

1. INTRODUCTION

The performance of most multi-view people detection algorithms depends on the quality of the extracted foreground silhouettes. The latter are the backbone of more advanced systems to detect, track and analyze people behavior [1]. Foreground silhouettes are binary masks representing the connected pixels belonging to the foreground of the scene. Static cameras can model the background of a scene by collecting statistics on every pixels. Porikli presents some of the methods in a brief survey [2]. Typically, a decade ago, Stauffer and Grimson modeled each pixel as a mixture of Gaussian with an on-line approximation for the update [3]. A lot of efforts have been dedicated to enhance these algorithms to best model the temporal intensity variation [4, 5]. However, sudden changes of lighting conditions dramatically affect the performance of the extraction process. The latters are only robust to gradual changes of lighting conditions whereas sudden changes such as turning on and off the indoor lights, spot light effects in exhibitions, camera flashes are not well addressed. Figure 1 illustrates the problem of any intensity-based background subtraction algorithm to extract foreground silhouettes when a sudden change occurs. Recent work have addressed the sudden change of illumination issue

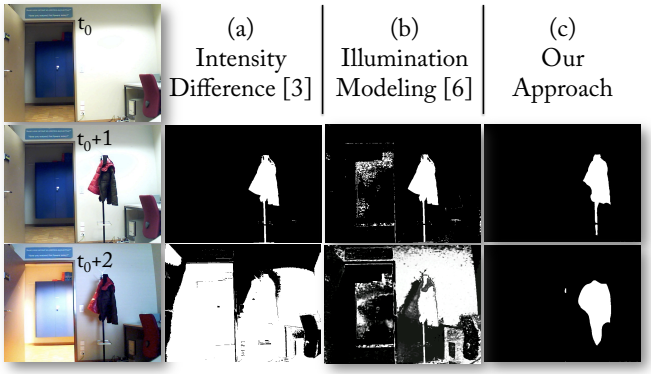
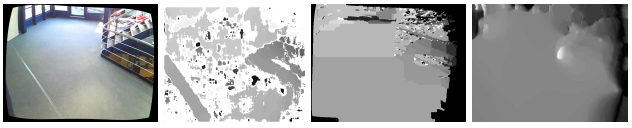


Fig. 1: Illustration of our TV disparity-based foreground extraction algorithm (right column) compared to a traditional intensity-based algorithm (2^{nd} and 3^{rd} column). The background is modeled at time t_0. We turned the light off at frame $t_0 + 2$ and succeed to locate the foreground object although two opposite intensity variation occurred (turning down the light in the room and up in the corridor).

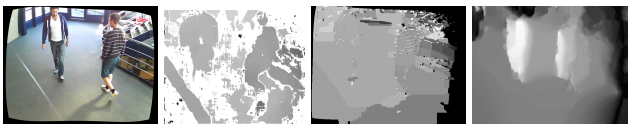
[6, 7] but either suppose that the change is global or do not work in environments when videos are projected in the background. As a result, we propose to tackle this problem with stereo imaging. Instead of classifying the image intensity as background or foreground, the estimated disparity map is processed to identify foreground regions.

There have been many attempts to do segmentation using stereo cameras, mainly for background subtraction applications [8, 9]. However, stereo vision algorithms are only used as a post processing tool to enhance the foreground silhouettes extracted given the intensity variation of each camera. Therefore, these algorithms fail again when sudden intensity changes occur.

Two strategies can been exploited to extract foreground silhouettes without considering the temporal intensity variation: (i) Temporal disparity variation , and (ii) stereo mismatch, where the disparity mismatch is calculated from an estimated background disparity [10]. The first approach depends on the performance of the disparity estimation algorithms. Although a lot of effort has been dedicated to find accurate algorithms to estimate disparity maps, the performances are often poor and not consistent across scenes. Algorithms often perform poorly on uniform and non-textured

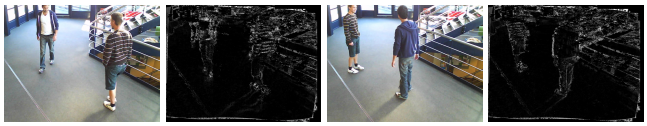


(a) Frame t_0 (b) Block Matching [11] (c) Graph Cuts [12] (d) TV framework [13]



(a) Frame t_1 (b) Block Matching [11] (c) Graph Cuts [12] (d) TV framework [13]

Fig. 2: Output of some well known algorithms to estimate disparity maps.



(a) Frame t_1 (b) Disparity mismatch at t_1 (c) Frame t_2 (d) Disparity mismatch at t_2

Fig. 3: Foreground silhouettes obtained with disparity mismatch.

regions (See Figure 2). Stereo mismatch algorithms compute the background disparity only. Even if the background disparity is correctly estimated, the extracted foreground silhouettes are still very noisy and sketchy (see Section 2). In this paper we propose to use a Total Variation (TV) framework for the extraction of dense foreground disparities to fill the sketchy look of the stereo mismatch and hence extract foreground silhouettes regardless the temporal changes in background appearance.

2. RELATED WORK: FOREGROUND EXTRACTION WITH STEREO MISMATCH

Disparity estimation at every time frame is computationally expensive. Advanced high complexity algorithms cannot be used for real-time segmentation. Conversely, stereo mismatch is a real-time alternative often proposed in the literature [8, 10, 14]. In such a case, the background disparity is computed once at the beginning. Then for every frame, the background disparity is used to wrap every right image on the left one to measure similarity. The foreground silhouettes are computed as follows:

$$F(t) = \begin{cases} 1 & \text{if } |I_L(t) - I_W(t)| > \theta, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where

$$I_W(t) = I_R(\vec{x} + D(0), t), \quad (2)$$

with $I_L(t)$ and $I_R(t)$ being respectively the left and right images of the stereo camera at time t. The wrapping operation, $I_R(\vec{x} + D(0), t)$, maps the image plane of the right image on the left one. By definition, if the background disparity map $D(0)$ is perfect, we have:

$$I_L(0) - I_W(0) = I_L(0) - I_R(\vec{x} + D(0), 0) = 0. \quad (3)$$

Equation 3 is equivalent to the brightness consistency assumption. To compute the disparity maps, we suppose that the brightness does not change between the left and right images. This can be done without loss of generality by supposing identical calibration or applying histogram matching methods.

Once foreground objects are present in the scene, the disparity $D(0)$ does not correct anymore the mismatch over the objects' region. Their silhouettes are correctly extracted if the objects have non-uniform color distribution. However, upon a uniform intensity, such approach extracts the contour of the foreground silhouettes instead of filled shapes, as required in further detection and tracking steps [1] (see Figure 3). Although pixels are not mapped correctly, they are compared with neighboring pixels having similar intensity value. Objects with uniform regions induce intensity differences on boundaries only. A solution would consist in filling these contours, but as they are not stable and almost never closed, some heavy and slow processing should be applied. In addition, stereo mismatch method in practice induces a lot of noise due to poor calibration. Therefore, we propose a Total Variation (TV) disparity estimation framework to extract foreground silhouettes based on the noisy stereo mismatch and hence promote filled silhouettes.

3. TOTAL VARIATION DISPARITY-BASED FOREGROUND EXTRACTION

We formulate the foreground extraction problem as a correspondence problem from the left image I_L to a wrapped one, I_W. The latter is obtained by applying the background disparity to the right image as in Equation 2. The correspondence problem is solved using a TV-ℓ_1 framework where the ℓ_1 norm penalize deviations from the brightness consistency assumption and a total variation regularization term penalize a sketch-like solution with contours and promote a solution with filled shapes. In other words, our approach to improve the foreground silhouette extraction is to replace the subtraction operator in the disparity mismatch by a TV-ℓ_1 disparity computation.

We first compute the background disparity between the left and right image, $D(0)$. Then, for every frame, the foreground image, $F(t)$, is computed by measuring the disparity between the wrapped and the left image. To get a mask, a simple threshold is used on the output.

Here is the complete algorithm:

1. We first compute the background disparity $D(0)$. Any algorithm can be used. Since it is an off-line process,

there is no constraint on the computational complexity. We use the TV ℓ_1 framework proposed in [13] to compute the disparity background.

2. For each new frame, we wrap the right image given the background disparity :

$$I_W = I_R(\vec{x} + D(0)) \quad (4)$$

3. The foreground disparity D_F is computed using the wrapped image I_W and the left image:

$$D_F = arg\min_{D_F} \sum_{\mathbf{x}} |I_L(\vec{x}) - I_W(\vec{x} + D_F)| + \lambda |\nabla D_F|, \quad (5)$$

where λ is the regularization parameter, and ∇ represents the discrete gradient operator. The fidelity term $|I_L(\vec{x}) - I_W(\vec{x} + D_F)|$ is similar to the stereo mismatch Equation 1 with the difference of the term D_F. As mentioned previously, in the presence of new objects, the background disparity will not be enough to compensate the stereo mismatch and hence need to be rectified by D_F. To solve Equation 5, we use the same multi-resolution GPU accelerated real-time algorithm proposed in [15]. The first step of the algorithm consist in a first order linearization of Equation 5:

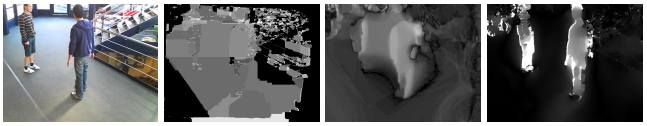
$$D_F = arg\min_{D_F} \sum_{\mathbf{x}} |I_L(\vec{x}) - I_W(\vec{x}) - < \nabla I_W(\vec{x}), D_F > | + \lambda \cdot |\nabla D_F|, \quad (6)$$

where $< ., . >$ is a scalar product. Then Equation 6 is solved with a first order primal-dual algorithmn consisting in an alternation of projection steps. For the details of the algorithm we refer to [15]. It is important to point out that since it is a multi-resolution scheme it is able to compensate even large disparities. Furthermore the number of levels used in the algorithm can be adjusted according to the maximum disparity to be estimated. In our experiments we found that usually few levels are enough to achieve good performances. Figure 5 illustrates the disparity D_F obtained as a solution of Equation 6.

4. Finally, the solution of Equation 6 needs to be thresholded. Ideally, any non-zero values should be considered as foreground. However, since a residual shift exists after the minimization process, we set a small threshold β (estimated heuristically).

4. EXPERIMENTAL RESULTS

We have evaluated our approach on three sequences: a scene made of two people walking (Figure 2 to 5), a conference room (Figure 6 and 7), and an indoor office (Figure 1). First, we have extracted foreground silhouettes given the temporal disparity variation, *i.e.* comparing the background disparity with the current estimated disparity. Figure 4 presents the silhouettes extracted given such approach. It can be seen that the estimated disparities are too noisy and not consistent across time. Although the graph cuts and TV approach are only illustrated in Figure 4, other local and global algorithms have also been tested and led to similar results. As mentioned previously, the performance of depth estimation algorithms depends on the content of the scene. The background disparity, $D(0)$, corresponds to an empty scene with poorly textured and uniform regions, which is one of the most challenging scenes for a disparity estimation algorithm. Therefore, the resulting disparity is noisy and leads to useless foreground silhouettes as illustrated in Figure 4. In contrast, using our TV disparity based approach, two people walking in the scene are correctly segmented as foreground (see Figure 5).

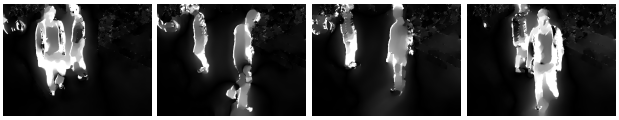


(a) Original image (b) Graph Cuts [12] (c) TV framework [13] (d) Our approach

Fig. 4: Foreground silhouettes obtained by temporal depth variation using two depth estimation techniques. For comparison purposes, we also present in (d) the output of our proposed TV disparity-based approach.

The TV framework clearly promotes filled shapes as opposed to the stereo mismatch techniques. In fact, the total variation regularization favors piece-wise smooth candidates leading to foreground silhouettes suitable for further processing step to understand a scene. Figure 6 illustrates the difference between the intensity-based approaches, the stereo mismatch approach presented in Section 2, and our proposed approach. Even if part of the scene has abrupt change of illumination, the foreground silhouette is still correctly segmented in Figure 6. Any abrupt change of the background appearance can happen as long as the the change does not affect the 3D modeling of the scene. Figure 7 shows the robustness of our algorithm when video is projected in the background.

The drawback of the proposed framework is its sensitivity to the parameters. Like in all variational optimizations, the convergence of the TV based disparity estimation is very sensitive to the lagrangian parameter λ in Equation 5, and on the appropriate choice of the number of levels in the multi-resolution scheme. In this paper we choose experimentally these parameters, in future works we plan to address the problem of their automatic selection. In addition, note that the foreground silhouettes are highly approximated but validate the requirement for dictionary-based people detection algorithms similar to the one presented in [1]. Future work will also investigate a more precise extraction required by other applications.



(a) Frame #100 (b) Frame #200 (c) Frame #300 (d) Frame #400

Fig. 5: Illustration of the output of our proposed TV disparity based foreground extraction algorithm before the thresholding step when two people are walking.

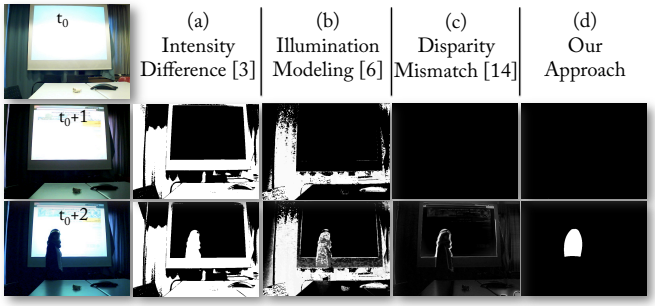


Fig. 6: Illustration of the output of our proposed TV disparity based foreground extraction algorithm where part of the scene has sudden change of illumination.

5. CONCLUSIONS

We have presented a TV disparity-based approach to extract foreground silhouettes with stereo cameras in a manner robust to sudden changes of background appearance. The TV framework penalizes the sketchy-like solution made of contours only and hence promotes filled silhouettes as required in many applications.

6. REFERENCES

[1] Alexandre Alahi, Laurent Jacques, Yannick Boursier, and Pierre Vandergheynst, "Sparsity driven people localization with a heterogeneous network of cameras," *Journal of Mathematical Imaging and Vision*, pp. 1–20, 2011.

[2] F. Porikli, "Achieving real-time object detection and tracking under extreme conditions," *Journal of Real-Time Image Processing*, vol. 1, no. 1, pp. 33–40, 2006.

[3] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *CVPR*, 1999.

[4] L. Tessens, M. Morbée, W. Philips, R. Kleihorst, and H. Aghajan, "Efficient approximate foreground detection for low-resource devices," in *ICDSC*, 2009.

[5] V. Cevher, A. Sankaranarayanan, M. Duarte, D. Reddy, R. Baraniuk, and R. Chellappa, "Compressive sensing for background subtraction," *ECCV*, pp. 155–168, 2008.

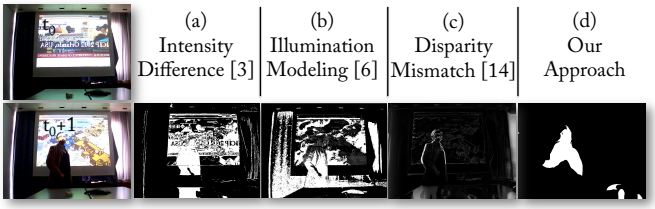


Fig. 7: Illustration of the output of our proposed TV disparity based foreground extraction algorithm where part of the scene has sudden change of illumination.

[6] J. Pilet, C. Strecha, and P. Fua, "Making background subtraction robust to sudden illumination changes," *ECCV*, pp. 567–580, 2008.

[7] G. Xue, J. Sun, and L. Song, "Background subtraction based on phase and distance transform under sudden illumination change," in *ICIP 2010*, pp. 3465–3468.

[8] S.N. Lim, A. Mittal, L.S. Davis, and N. Paragios, "Fast illumination-invariant background subtraction using two views: Error analysis, sensor placement and applications," in *CVPR 2005*, pp. 1071–1078.

[9] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother, "Bi-layer segmentation of binocular stereo video," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2005, pp. 407–414.

[10] Yuri A. Ivanov, Aaron F. Bobick, and John Liu, "Fast lighting independent background subtraction," *International Journal of Computer Vision*, 2000.

[11] Karsten Mühlmann, Dennis Maier, Jürgen Hesser, and Reinhard Männer, "Calculating dense disparity maps from color stereo images, an efficient implementation," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 79–88, 2002.

[12] Yuri Boykov, Olga Veksler, and Ramin Zabih, "Fast approximate energy minimization via graph cuts," in *ICCV*, 1999, pp. 377–384.

[13] L. Bagnato, P. Frossard, and P. Vandergheynst, "A variational framework for structure from motion in omnidirectional image sequences," *Journal of Mathematical Imaging and Vision*, pp. 1–12, 2009.

[14] Wei Sun and Stephen P. Spackman, "Multi-object segmentation by stereo mismatch," *Mach. Vis. Appl.*, vol. 20, no. 6, pp. 339–352, 2009.

[15] Andreas Weishaupt, Luigi Bagnato, and Pierre Vandergheynst, "Fast Structure from Motion for Planar Image Sequences," in *Eusipco*, 2010.