Bilayer Segmentation Augmented with Future Evidence

Silvio Ricardo Rodrigues Sanches[1], Valdinei Freire da Silva[2], and Romero Tori[1]

[1] Escola Politécnica da Universidade de São Paulo, Brasil, Av Professor Luciano Gualberto, Trav. 3, 158, Cidade Universitária, 05508-900, São Paulo-SP, Brasil
silviorrs@usp.br,tori@acm.org

[2] Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, Av. Arlindo Béttio, 1000, Ermelino Matarazzo, 03828-000, São Paulo-SP, Brasil
valdinei.freire@usp.br

Abstract. This paper presents an algorithm that augments a previous model known in the literature for the automatic segmentation of monocular videos into foreground and background layers. The original model fuses visual cues such as color, contrast, motion and spatial priors within a Conditional Random Field. Our augmented model makes use of bidirectional motion priors by exploiting future evidence. Although our augmented model processes more data, it does so with the same time performance of the original model. We evaluate the augmented model within ground truth data and the results show that the augmented model produces better segmentation.

Keywords: bilayer segmentation, computer vision, image understanding.

1 Introduction

The image segmentation problem (the extraction of elements of interest in images or videos) has been under research since the beginning of the last century with the industry [24]. The industry of film and television productions traditionally use methods which extract one or more elements from an image or a video frame – most of these elements are people in the foreground – to create scenes from the combination of them with a new background [20,21]. Until the late 1970s such extraction was based on optical analog technology [5].

Traditional methods for image segmentation assume that the video frame was captured in a controlled environment, with a single color as background (usually blue or green) and with the environment lights configured to keep that color uniform [6,14,21]. By assuming a known background color, segmentation can be done in real-time and with low error.

Since the 1980s, new methods based on digital technology [6] have been developed and today there are methods able to extract elements not only in real-time but also from natural images (without a single color background) [2,4,18,22,26]. This possibility has boosted research in new areas of application, especially those

B. Murgante et al. (Eds.): ICCSA 2012, Part II, LNCS 7334, pp. 699–711, 2012.

in which the elements of interest are people in the foreground. Videoconferencing, videochat [4,18,26,25] and other systems [17] are examples of applications which can replace an original background before sending each video frame to remote users.

In this paper we address the efficient extraction of foreground layer (bilayer segmentation) to background substitution applications in which input streaming is captured in a natural environment by a monocular camera. The algorithm proposed here extends the segmentation model presented by Criminisi *et al.* [4]. The segmentation problem is modeled by a Conditional Random Field (CRF) which fuses color, temporal and spatial information. The temporal information is considered as a prior and it is conjugated with color and spatial models of observation in order to determine an *a posteriori* information. We improve on such a model by augmenting it with one future frame in the CRF model.

The rest of this paper is organized as follows: section 2 presents some previous researches related to bilayer segmentation and section 3 illustrates the methods which use temporal information. Section 4 introduces the basic model used in our solution, whereas the section 5 describes our augmented model. Section 6 discusses experimental results, and section 7 concludes the papper.

2 Related Work

Layer extraction [2,4,8,10,15,18,19,25] has long been an active area of research in computer vision. In recent approaches, a common characteristic is to treat the segmentation problem as an energy minimization problem. In a binary segmentation task each pixel of the processing frame is labeled as background or foreground (0 or 1), without considering fractional values to represent transparency. Briefly, from a set of pixels P and a set of labels L (in this case two labels), the goal is to find a label function $f : P \rightarrow L$ which minimizes a specific energy function [9][1].

In order to classify pixels, recent work in bilayer segmentation area has produced algorithms based on either depth [7,22] or motion [4,26]. Other algorithms require initialization in the form of a clean image of background [18].

Stereo-based segmentation [7] seems to achieve the most robust results for layer extraction with depth information. However, binocular video can be restrictive for some applications as well as approaches based in time-of-flight sensors for depth estimation [22]. In a teleconferencing or a videochat most users have only a single conventional web camera [18] and the necessity for calibration of two cameras for stereo is inconvenient [4].

Motion-based segmentation can be achieved by estimating optical flow [1], but, in the context of natural environments, the foreground motion cannot be described well by such rigid models. In addition, the optical flow computation is expensive [26].

[1] Determination of fractional transparency of pixel is necessary for precision in layer extraction. However, it can be determined after the binary segmentation, by techniques such as border matting [16].

Interactive color/contrast-based segmentation techniques have been demonstrated to be effective [3,16]. However, as demonstrated in [4], segmentation based on color/contrast alone is beyond the capability of fully automatic methods.

On the other hand, recent approaches show that fusing a variety of cues, for example, color, contrast and spatial priors [4,18,25,26] can produce segmentation computable in real-time with accuracy similar to the one obtained from stereo-based segmentation. Whereas some of those systems assume static background [4], others support distracting events and require no initialization [26].

In our model, visual cues such as color, contrast, motion and spatial priors are fused together within a CRF model, where the motion cue exploits bidirectional evidence in order to improve bilayer segmentation.

3 Temporal Information in Bilayer Segmentation Approaches

In applications where the bilayer segmentation is performed in controlled environments the misclassification of pixels can be avoided by user interventions. The environment light can be configured and the elements in the scene can be positioned so that the single background color remains constant. This can avoid the occurrence of shadows, reflections or noise on the background which are potential sources of error.

However, in natural environments, the background is arbitrary and any problematic situation must be handled by segmentation algorithm, avoiding user intervention. In such cases, as there is no prior knowledge about the background color, other information that can be obtained from the frame sequence become necessary, especially for segmentation methods based on monocular video.

Automatic and computationally efficient methods for real-time segmentation make use of these information as a set of *cuts* [4,26]. Color, contrast and motion are examples of cuts which are probabilistically combined by a framework of energy minimization.

As demonstrated in [4,18,25,26], information obtained from previous frames have proved important to aid the estimation current-frame labels. In those work, the temporal information is a cut in the energy minimization framework and it was used to identify pixels in movement. Due to the real-time restriction, the temporal information has been based only on the evidence from its past.

Although considering evidence from future (or bidirectional) may be prohibited in many applications, real-time ones in which some delay is already expected, the delay of one frame may be imperceptible. We demonstrate that this information is relevant for segmentation algorithms.

4 Notation and Basic Model

This section describes notation regarding frame observations and the probabilistic model for foreground/background segmentation proposed in [4] to which we refer as the basic model.

4.1 Notation

Given an input sequence of images, a frame is represented as a matrix

$$z = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,Y} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,Y} \\ \vdots & \vdots & \ddots & \vdots \\ z_{X,1} & z_{X,2} & \cdots & z_{X,Y} \end{bmatrix} \tag{1}$$

of pixels in the YUV color space. A frame at time t is denoted z^t. Temporal derivatives are denoted $\dot{z} = [\dot{z}_{x,y}]_{X \times Y}$ and are computed as

$$\dot{z}^t = |G(\mathbf{0}; \sigma_T) * z^t - G(\mathbf{0}; \sigma_T) * z^{t-1}| \tag{2}$$

at each time t, where $G(\mathbf{0}; \sigma_T)$ is a 2D centralized Gaussian kernel with standard deviation σ_T and $*$ is the convolution operator. Spatial gradients $g = [g_{x,y}]_{X \times Y}$ are computed by convolving the frames with first-order derivative of Gaussian kernels with standard deviation σ_S, i.e.,

$$g^t = \sqrt{\left(\frac{\partial G(\mathbf{0}; \sigma_S)}{\partial x} * z^t\right)^2 + \left(\frac{\partial G(\mathbf{0}; \sigma_S)}{\partial y} * z^t\right)^2} . \tag{3}$$

As in [4], we use $\sigma_S = \sigma_T = 0.8$. Spatio-temporal derivatives are computed on the Y channel only. Motion observation at time t is denoted $m^t = (g^t, \dot{z}^t)$. Given a sequence of image data $z^1, z^2, \ldots, z^t$ and a sequence of motion data $m^1, m^2, \ldots, m^t$, the segmentation task is to infer a binary label $\alpha^t_{x,y} \in \{F, B\}$ for every pixel in the current frame. F and B denote foreground and background, respectively.

4.2 Basic Model

The probabilistic model for layer extraction proposed in [4] uses an energy minimization framework and extends previous energy models for segmentation [3,7,16]. The model is a Conditional Random Field (CRF) [12] with independent terms that are set discriminatively, i.e., instead of working with join distributions, conditional distributions are considered [11]. The CRF models the conditional probability:

$$p(\alpha^1, \ldots, \alpha^t | z, \ldots, z^t, m^1, \ldots, m^t) \propto \exp - \left\{ \sum_{t'=1}^{t} E^{t'} \right\} \tag{4}$$

where $E^t = E(\alpha^t, \alpha^{t-1}, \alpha^{t-2}, z^t, m^t)$.

The energy E^t associated with time t is a sum of terms in which likelihood and prior are not entirely separated. The energy decomposes as a sum of four terms:

$$\begin{aligned} &E(\alpha^t, \alpha^{t-1}, \alpha^{t-2}, z^t, m^t) = \\ &\quad \eta V^T(\alpha^t, \alpha^{t-1}, \alpha^{t-2}) + \gamma V^S(\alpha^t, z^t) \\ &\quad +\rho U^C(\alpha^t, z) + \phi U^M(\alpha^t, \alpha^{t-1}, m^t), \end{aligned} \tag{5}$$

in which the first two terms are "prior-like" and the second two are observation likelihoods. η, γ, ρ and ϕ are normalizing parameters.

Temporal prior term $V^T(\cdot)$ imposes a tendency to temporal continuity of segmentation labels. Second-Order Markov chain is used in the energy minimization framework to incorporate the intuition that a pixel that was in the background at time $t-2$ and in the foreground at time $t-1$ is far more likely to remain in the foreground at time t than to go back to the background. The temporal transition priors are learned from labeled data. The temporal prior term is denoted

$$V^T(\alpha^t, \alpha^{t-1}, \alpha^{t-2}) = \sum_{m=1}^{X} \sum_{n=1}^{Y} [-\log p(\alpha^t_{x,y} | \alpha^{t-1}_{x,y}, \alpha^{t-2}_{x,y})]. \tag{6}$$

Spatial prior term $V^S(\cdot)$ is an Ising term, imposing a tendency to spatial continuity of labels, and the term is inhibited by high contrast. Let C be the set of pairs of neighboring pixels in a frame[2]; and z_i and α_i be the values of pixel i in the YUV color space and the binary label to be attributed to pixel i respectively, i.e., i indexes a pixel in the matrices z and α. The Ising term is represented by an energy of the form

$$V^S(\alpha, z) = \sum_{i,j \in C} [\alpha_i \neq \alpha_j] \left(\frac{\epsilon + e^{-\mu ||z_i - z_j||^2}}{1 + \epsilon} \right). \tag{7}$$

The contrast parameter μ is chosen to be $\mu = (2\langle ||z_i - z_j||^2 \rangle)^{-1}$, where $\langle \cdot \rangle$ denotes expectation over all pairs of neighbors in an image sample.

The energy term $V^S(\alpha, z)$ represents a combination of an Ising prior for labeling coherence together with a contrast likelihood that acts to discount partially the coherence terms. The constant ϵ is a "dilution" constant for contrast. We set $\epsilon = 1$ as it was done in [7].

Color likelihood term $U^C(\cdot)$ evaluates the evidence for pixel labels using the color distributions in foreground and background. Likelihoods are modeled as histograms in the YUV color space. This term is defined as:

$$U^C(\alpha, z) = -\sum_{m=1}^{X} \sum_{n=1}^{Y} \log p(z_{x,y} | \alpha_{x,y}). \tag{8}$$

In our experiments the foreground color likelihood model is learned from a first ground-truth segmented frame. The likelihoods are then stored in 3D look-up

[2] Here we work with a neighborhood of 4 neighbors, i.e., neighbors in each cardinal direction.

tables. The distribution is represented as a smoothed histogram to avoid overfitting within initialization.

Motion likelihood term $U^M(\cdot)$ uses spatial and temporal derivatives $m = (g, \dot{z})$ to capture the characteristics of the features under foreground and background conditions.

According to [4], the immediate history of the segmentation of a pixel falls into one of four classes: FF, BB, FB and BF. The observed image motion features $m^t_{x,y} = (g^t_{x,y}, \dot{z}^t_{x,y})$ at time t is conditioned on those combinations of the segmentation labels $\alpha^{t-1}_{x,y}$ and $\alpha^t_{x,y}$. Temporal derivative $\dot{z}^t_{x,y}$ is computed from frames $t-1$ and t, so it should depend on segmentations of those frames.

The motion likelihood is learned from some labeled ground-truth data and then stored as 2D histograms to use in likelihood evaluation. The likelihoods are evaluated as part of the total energy, in the term

$$U^M(\alpha^t, \alpha^{t-1}, m^t) = -\sum_{x=1}^{X}\sum_{y=1}^{Y} \log p(m^t_{x,y} | \alpha^t_{x,y}, \alpha^{t-1}_{x,y}). \tag{9}$$

4.3 Energy Minimization

Before minimizing energy E^t some considerations are done. First, parameters z^t and m^t are observations and, since they are extract from the current frame and the previous frame, they are not considered when minimizing energy E^t and we make it clear by writing $E^t = E(\alpha^t, \alpha^{t-1}, \alpha^{t-2} | z^t, m^t)$. Second, at time $t = 1$ parameters α^{t-1} and α^{t-2} are meaningless. Third, at time $t = 2$ parameter α^{t-2} is meaningless. Then, when labeling a frame at time $t = 1$, only terms V^S and U^C are meaningful and we have an energy $E^1 = E(\alpha^1 | z^t, m^t)$, and when labeling a frame at time $t = 2$, only terms V^S, U^C and U^M are meaningful and we have an energy $E^2 = E(\alpha^2, \alpha^1 | z^t, m^t)$.

Given the previous considerations, the labeling of pixels proceeds as follows:

- at time $t = 1$ minimizes energy $E^1 = E(\alpha^1 | z^t, m^t)$ and obtains $\hat{\alpha}^1$;
- by considering $\hat{\alpha}^1$ as an observation, at time $t = 2$ minimizes energy $E^2 = E(\alpha^2 | \hat{\alpha}^1, z^t, m^t)$; and
- by considering $\hat{\alpha}^{t-2}$ and $\hat{\alpha}^{t-1}$ as observations, at any time $t \geqslant 3$ minimizes energy $E^t = E(\alpha^t | \hat{\alpha}^{t-1}, \hat{\alpha}^{t-2}, z^t, m^t)$.

The simplification of considering $\hat{\alpha}^{t-2}$ and $\hat{\alpha}^{t-1}$ as observations decreases the neighborhood of a pixel when minimizing the energy function. Temporal dependences are solved by doing so and only spatial dependences are need to be solved. Since the energy E^t models a CRF, we can describe it as [11]:

$$E^t = \sum_{i \in \mathcal{S}} \left[A_i(\alpha^t_i, \mathbf{o}^t) + \sum_{j \in \mathcal{N}_i} I_{ij}(\alpha^t_i, \alpha^t_j, \mathbf{o}^t) \right],$$

where $\mathcal{S}$ is the set of pixels in a frame, $\mathbf{o} = (\hat{\alpha}^{t-1}, \hat{\alpha}^{t-2}, z^t, m^t)$ is the observation at time t, $\mathcal{N}_i$ is the neighborhood of pixel i, A_i is the association potential and I_{ij} is the interaction potential.

Finally, to minimize energy we considered the implementation of graph cut done by Kolmogorov and Zabih [9].

5 Augmenting the Basic Model

Although Criminisi *et al.* [4] make use of motion characteristics in the motion term U^M, Criminisi *et al.* make use of motion characteristics only related to past frames. Since the labels α^t influences temporal derivatives, we can consider such derivatives backwards and forwards in time. In [4] only the first derivative is explored. Although considering evidence from future may be prohibited in many applications, if some delay is already expected in some real-time applications, the user would not perceive the delay of one more frame. Fig. 1 shows the relation between each variable considered in the CRF.

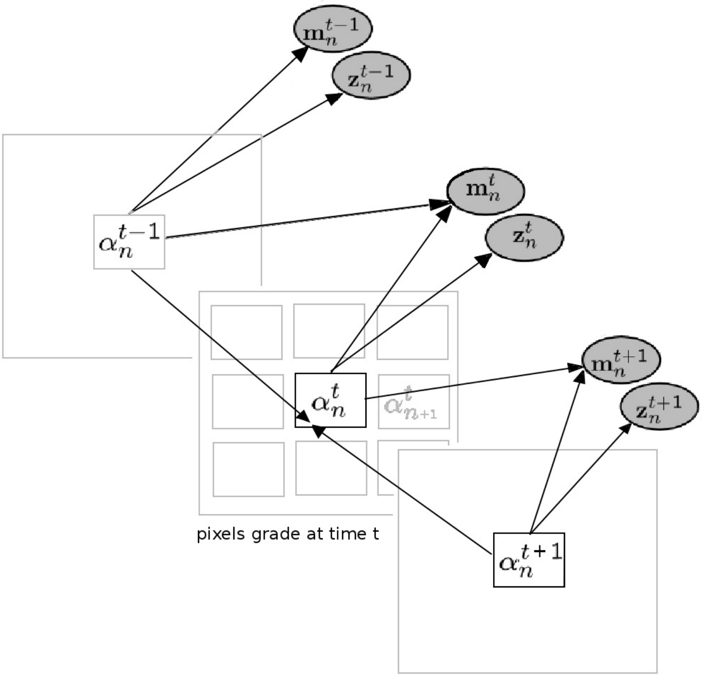

Fig. 1. Augmented CRF Model

The model by Criminisi *et al.* defines the observation $m^t = (g^t, \dot{z}^t)$ and minimizes the likelihood $p(g^t, \dot{z}^t|\alpha^t, \alpha^{t-1})$. However, if we wait for one more frame, another evidence $\dot{z}^{t+1}$ can be observed, given the opportunity of minimizing the likelihood $p(g^t, \dot{z}^t, \dot{z}^{t+1}|\alpha^t, \alpha^{t-1})$. If the delay in one frame is acceptable, such evidence may improve the segmentation of the foreground layer.

In order to evaluate the value of such evidence, we calculate the entropy when labeling sequence of images in videos which simulates videochat interaction. Given observations $\alpha^{t-1}, \dot{z}^{t}, \dot{z}^{t+1}$ and g^{t}, we combined them to classify labels α^{t}. In order to test the influence of each one of the observations, we grouped them into four groups and classified labels α^{t} with the maximum likelihood criteria. The likelihood was obtained from a grounded dataset and tested against itself. Table 1 shows our results: entropy and error rate for each group. Combining only two derivatives evidences ($\dot{z}^{t}, \dot{z}^{t+1}$ or g^{t}) gives similar values, but combining all of them together increases the classification based only on motion significantly if observation α^{t-1} is correct. Note that we cannot obtain the same results from table 1, since in our framework α^{t-1} is classified in previous steps and may present errors.

Table 1. Entropy Analysis

	Entropy	Error Rate
$g^{t}, \dot{z}^{t}, \alpha^{t-1}$	0.2100	0.033
$\dot{z}^{t}, \dot{z}^{t+1}, \alpha^{t-1}$	0.2013	0.031
$g^{t}, \dot{z}^{t+1}, \alpha^{t-1}$	0.2628	0.045
$g^{t}, \dot{z}^{t}, \dot{z}^{t+1}, \alpha^{t-1}$	0.1252	0.017

Since $p(g_n^{t}, \dot{z}_n^{t}, \dot{z}_n^{t+1} | \alpha_n^{t}, \alpha^{t-1})$ is kept in a look-up table and the future temporal derivative $\dot{z}_n^{t+1}$ should be calculated anyway at time $t+1$, our augmented algorithm calculates it in advance and no difference in computational time is observed. However, the size of the look-up table is multiplied by the spectrum of $\dot{z}^{t+1}$, increasing considerably the size of the look-up table.

If the size of the table is larger, we must also collect more data in order to learn the conditional probabilities $p(g_n^{t}, \dot{z}_n^{t}, \dot{z}_n^{t+1} | \alpha_n^{t}, \alpha^{t-1})$. Again, we analyze labeled videos in order to determine a better use of the look-up table, decreasing its size and improving its generalization when learning. Fig. 2 shows the histogram for the variable $\dot{z}$. It is clear the concentration around small values and the sparseness when values get large. Analyzing g proved to be similar.

Instead of using uniform discretization of the derivative spectrum, following our analysis, we divide the spectrum of derivatives in 39 bins. Around zero, bins are smaller, whereas bins increases as they are distant from zero. The size of bins from zero values to larger one are: five bins with size 1, six bins with size 2, twelve bins with size 4, eight bins with size 8, and eight bins with size 16.

6 Experiments

In order to experiment our ideas we compared both Criminisi *et al.* algorithm and our augmented version. We used a database of video images with 38 labeled video sequences which are available in [13]. In the experiments we derive our augmented version of Criminisi *et al.* from [23].

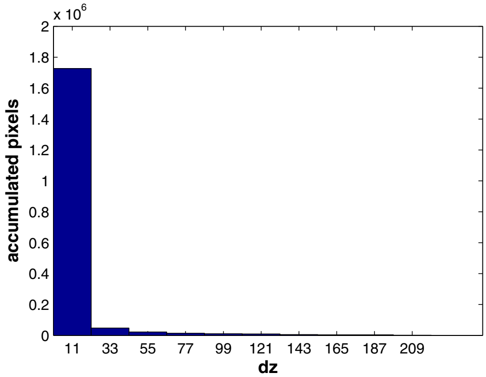

Fig. 2. Distribution of $\dot{z}^t$ values for a video

Since frames are labeled from 10-to-10 or 5-to-5 frames, we opted to use only such frames in our experiments. The labels of the first frames of each video is considered to be known in order to initialize the color likelihood model as well as initializing the prior regarding foreground layer. First, we obtained the temporal prior and the motion likelihood within all of the 38 labeled video sequences. In order to tune the parameters of segmentation models, we sampled randomly a hundred tuple of parameters $\langle \eta, \gamma, \rho, \phi \rangle$ and chose the tuple that obtained the best evaluation. Since the absolute values of parameters are irrelevant, we normalized parameter values by fixing $\gamma = 1$ and we sorted η, ρ and ϕ from interval $(0.0, 0.2)$ uniformly. Because γ proved to be more relevant in preliminary experiments, we sorted the parameters η, ρ and ϕ smaller than the parameter γ.

The evaluation of tuples was obtained by calculating the mean error among frames in every sequence. The error per frame was simply calculated as the error rate when compared to ground truth data, i.e., at frame t, given the ground truth α^t and the estimated label $\hat{\alpha}^t$, the error rate at frame t is given by

$$\epsilon^t = \epsilon(\alpha^t, \hat{\alpha}^t) = \frac{\sum_{x=1}^{X} \sum_{y=1}^{Y} |\alpha^t - \hat{\alpha}^t|}{XY}.$$

Remember that labels are binary, then $|\alpha^t - \hat{\alpha}^t| \in \{1, 0\}$. By considering all of the 38 video sequences we obtained the parameters $\eta = 0.0018$, $\gamma = 1$, $\rho = 0.0338$, $\phi = 0.0413$.

Because of our small database of video sequences, instead of using the same parameters to evaluate all of the video sequence, we opted by using leave-one-out method to training/testing video sequences. When evaluating the performance of each method regarding a video sequence, we took it out of the set of samples and chose the best parameters for the other 37 video sequences. In Fig. 3 foreground and background of a video sequence were separated automatically by our augmented model.

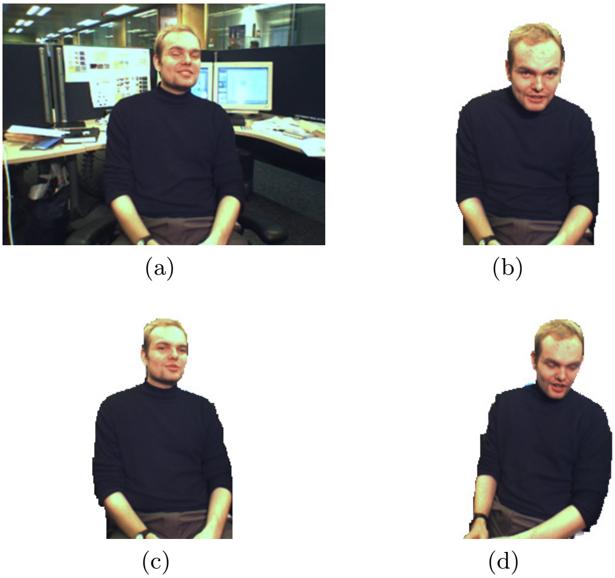

Fig. 3. Binary Segmentation. (a) A frame from the MS test sequence. (b,c and d) Automatic foreground extraction results for three frames by our augmented method.

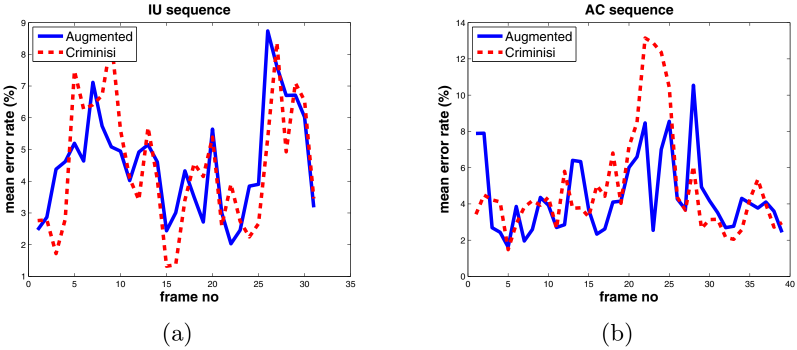

Fig. 4. Comparison between Criminisi *et al.* method and our augmented method. (a) IU test sequence and (b) AC test sequence. Note that the model parameters η, γ, ρ e ϕ were fully optimized for best performance.

We used the labeled video sequences for experimenting our augmented version against Criminisi *et al.* algorithm. Our approach had a mean error rate through the whole video of 0.041 whereas the basic approach presented a mean error rate of 0.054. Despite our small database, by applying t-student test we observed that our method was better than the original one with a significance of 0.18. Fig. 4 shows our results in IU and AC test sequences[3]. We found that the IU sequence

[3] The name of sequences IU, AC and MS are given in [13].

exhibit high illumination variation. Our best relative result were observed in this test sequence.

As a final example, Fig. 5 shows the results of our augmented method on a frame sequence. The original background was replaced with a new one. In this demonstration, the colour likelihood model was initialized manually.

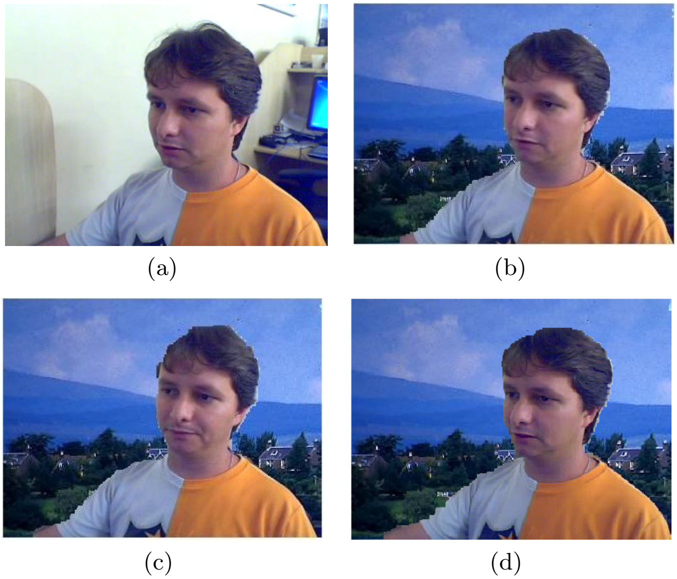
(a)
(b)
(c)
(d)

Fig. 5. A final example of Binary Segmentation and Background Substitution. (a) A frame with original background; (b, c and d) automatic background substitution for several frames by our augmented method. Note that no method for transparency of pixel was applied.

7 Conclusion

This paper has addressed the problem of bilayer segmentation of monocular video sequences. We extend the model presented in [4] to improve segmentation without incurring in more computational time. We accomplish it by extending the spatio-temporal coherence considering bidirectional evidence from a frame without losing real-time characteristic.

Although the delay of one frame can be prohibited in some high-performance application, most interactive applications accept small delays. If a high frame rate is considered, this delay can even be imperceptible.

The results show that an improvement was obtained, however more specific experiments must be done in order to detect in which kind of situation it is fruitful considering future frames.

Acknowledgment. Silvio Sanches would like to thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – scolarship) and Valdinei Silva would like to thank FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo – proc. 11/19280-8) for their financial supports. The video sequences and the ground truth used in this research are courtesy of the Microsoft Corporation (Free research data [13]).

References

1. Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. International Journal of Computer Vision 12, 43–77 (1994)
2. Bergen, J., Burt, P., Hingorani, R., Peleg, S.: A three-frame algorithm for estimating two-component image motion. IEEE Transactions on Pattern Analysis and Machine Intelligence 14(9), 886–896 (1992)
3. Boykov, Y.Y., Jolly, M.P.: Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In: IEEE International Conference on Computer Vision, vol. 1, pp. 105–112. IEEE Computer Society, Los Alamitos (2001)
4. Criminisi, A., Cross, G., Blake, A., Kolmogorov, V.: Bilayer segmentation of live video. In: CVPR 2006: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 53–60. IEEE Computer Society, Washington, DC (2006)
5. Foster, J.: Mattes and Compositing Defined. In: The Green Screen Handbook: Real-World Production Techniques, pp. 3–15. John Wiley and Sons Ltd., Chichester (2010)
6. Gibbs, S., Arapis, C., Breiteneder, C., Lalioti, V., Mostafawy, S., Speier, J.: Virtual studios: an overview. IEEE Multimedia 5(1), 18–35 (1998)
7. Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., Rother, C.: Bi-layer segmentation of binocular stereo video. In: CVPR 2005: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 407–414. IEEE Computer Society, Washington, DC (2005)
8. Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., Rother, C.: Probabilistic fusion of stereo with color and contrast for bilayer segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(9), 1480–1492 (2006)
9. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? IEEE Transactions on Pattern Analysis and Machine Intelligence 26(2), 147–159 (2004)
10. Kumar, M., Torr, P., Zisserman, A.: Learning layered motion segmentations of video. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 1, pp. 33–40 (2005)
11. Kumar, S., Hebert, M.: Discriminative random fields: A discriminative framework for contextual interaction in classification. In: ICCV 2003: Proceedings of the Ninth IEEE International Conference on Computer Vision, p. 1150. IEEE Computer Society, Washington, DC (2003)
12. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML 2001: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann Publishers Inc., San Francisco (2001)
13. Microsoft research – free research data, `http://research.microsoft.com/en-us/projects/i2i/data.aspx`

14. Mishima, Y.: Soft edge chroma-key generation based upon hexoctahedral color space. U.S. Patent 5,355,174 (1994)
15. Parolin, A., Fickel, G.P., Jung, C.R., Malzbender, T., Samadani, R.: Bilayer video segmentation for videoconferencing applications. In: 2011 IEEE International Conference on Multimedia and Expo. (ICME), pp. 1–6 (2011)
16. Rother, C., Kolmogorov, V., Blake, A.: "Grabcut": interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23(3), 309–314 (2004)
17. Sanches, S.R.R., Tokunaga, D.M., Silva, V.F., Sementille, A.C., Tori, R.: Mutual occlusion between real and virtual elements in augmented reality based on fiducial markers. In: 2012 IEEE Workshop on Applications of Computer Vision (WACV), pp. 49–54 (2012)
18. Sun, J., Zhang, W., Tang, X., Shum, H.Y.: Background Cut. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 628–641. Springer, Heidelberg (2006)
19. Torr, P.H.S., Szeliski, R., Anandan, P.: An integrated bayesian approach to layer extraction from image sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(3), 297–303 (2001)
20. Vlahos, P.: Composite photography utilizing sodium vapor illumination. U.S. Patent 3,095,304 (1963)
21. Vlahos, P.: Comprehensive electronic compositing system. U.S. Patent 4,100,569 (1978)
22. Wang, L., Zhang, C., Yang, R., Zhang, C.: Tofcut: Towards robust real-time foreground extraction using a time-of-flight camera. In: Fifth International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT) (2010)
23. Implementation of "bilayer segmentation of live video", `http://vision.caltech.edu/projects/yiw/FgBgSegmentation`
24. Williams, F.D.: Method of taking motion pictures. U.S. Patent 1,273,435 (1918)
25. Yin, P., Criminisi, A., Winn, J., Essa, I.: Tree-based classifiers for bilayer video segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8. IEEE Computer Society, Los Alamitos (2007)
26. Yin, P., Criminisi, A., Winn, J., Essa, I.: Bilayer segmentation of webcam videos using tree-based classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(1), 30–42 (2011)