Improvement of background subtraction by mask constraints

Vadim Konushin[1,2], Anton Konushin[1]
[1]Graphics and Media Lab, Moscow State Lomonosov University, Moscow, Russia
[2] Keldysh Institute of Applied Mathematics, Russian Academy of Sciences
E-mail: {vadim, ktosh}@graphics.cs.msu.ru

Abstract

In this paper we propose an improvement for background subtraction algorithms for specific video surveillance scenario. We consider a case, when a video camera is attached to a wall and observes people walking by or coming up to the camera. We propose 2 foreground mask models and show how to integrate these new mask constraints into common background subtraction techniques – pixel-based algorithms and methods, based on graphical models.

The proposed modifications have no parameters and add little computational overhead. Experiments, conducted on our own video sequences, demonstrate segmentation accuracy improvement of the modified algorithms.

Keywords: *Foreground extraction, Background subtraction, Markov Random Field, GraphCut.*

1. INTRODUCTION

Human-computer interfaces, based on video surveillance, are a rapidly developing field of research and have a wide range of applications, e.g. smart advertisement. Such systems can react differently based on subject's estimated relative position, gaze direction, gender, height, age, race, gestures, etc. However, an accurate estimation of these parameters are required, otherwise reaction can be awkward and confusing.

Foreground extraction is an important preprocessing step of most operations in video-based human interfaces. Many detection and tracking algorithms detect foreground blobs first and then match them to the tracked objects. Person classification and recognition algorithms could benefit from an accurate foreground mask by using the context: hair color, clothes texture, person height could bring the additional information for a more accurate classification in comparison with a common face classification.

Most foreground extraction techniques are designed for a general video surveillance scenario. In this work we concentrate on a single video surveillance scenario, when a camera is attached to a wall at 1-2m height and observes people coming up to it, see Fig. 1 (a). This is a common case for smart advertisement systems, smart kiosks, etc. For this particular surveillance scenario we propose two different foreground mask models. These models constraint background subtraction: by allowing a smaller set of possible masks, they make the whole algorithm more stable for this scenario. We show how to modify standard pixel-based algorithms, and algorithms, based on Markov Random Fields to account for this model.

The rest of the paper is organized as follows: section 2 describes current background subtraction techniques, in section 3 we define the proposed masks models. Modifications of standard background subtraction methods are given in section 4. In section 5 we report the obtained results. Section 6 concludes the paper.

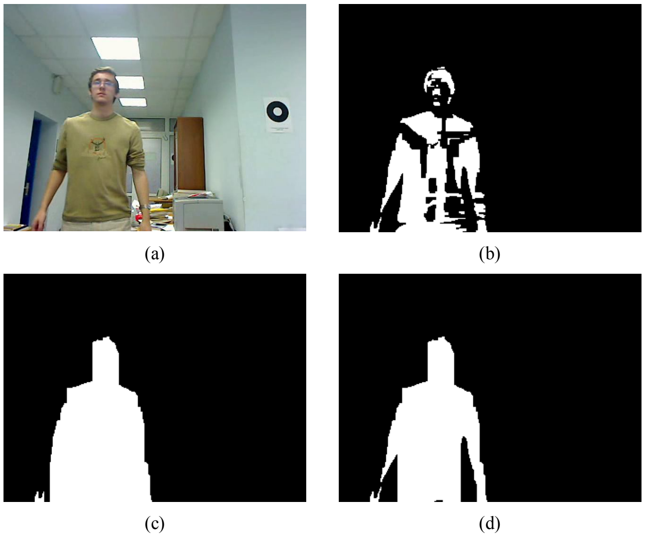

Figure 1: Example of background subtraction (a) a source frame, (b) result by an algorithm, based on Markov Random Field, (c) result of algorithm's modification, Mask Model 1, (d) result of algorithm's modification, Mask Model 2. See text for the details.

2. RELATED WORK

We can coarsely classify most background subtraction techniques into 3 classes: pixel-based [1], [2], [3], patch-based [4] and the algorithms that are based on some graphical models – Markov Random Field (MRF) and Conditional Random Field (CRF) [5], [6], [7].

Pixel-based algorithms model background color distributions for each pixel independently and for every video frame decide, whether the pixel is occluded by comparing it with a stored background model. Common examples are to model pixel color distribution with a single gaussian [1], gaussian mixture model [2] or non-parametrically [3].

Patch-based algorithms [4] model distribution of image patches. By using patches instead of single pixels, they can consider image texture in addition to pixel color, so they can be more robust to noise and in some cases to illumination change. But again, the classification into foreground/background is made independently of other patches.

The last group of the algorithms treats the whole image as a graphical model [5], [6], [7]. They fuse different features (colour, contrast, motion) by means of MRF or CRF. Then maximum a posteriori (MAP) solution is found with the use of GraphCut [8].

Most algorithms don't impose any constraints on the resulting foregound mask, so they can be applied to any video surveillance scenario, though most tests consider the security applications: camera watching a parking, a shop, an entry to some office.

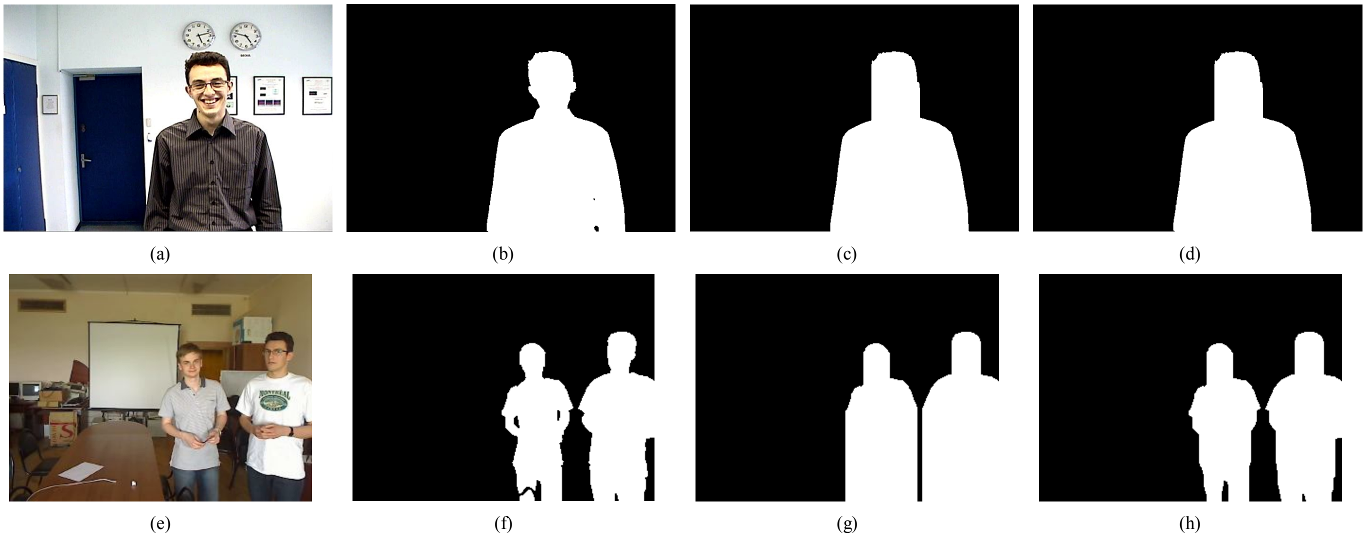

Figure 2: Mask models example. (a, e) – source frames, (b, f) – ground truth foreground maks, (c, g) –mask model 1, (d, h) – mask model 2.

Some mask constraints can be used only in a postproccessing: for choosing a radius of morphology operations or classification of resulting blobs into the objects of different classes or noise based on their size.

Some CRF-based methods [6], [7] specifically consider the case of video-chat type sequences as they are designed for possible background substitution during video conferences. They learn motion cues to differentiate a closest person from people, walking behind him during the video conference.

In this paper, we focus on a different scenario. A video-camera is attached to a wall or some stand, approximately on a human's height and records people walking by or coming up to the camera.

We propose 2 types of mask models for this scenario. We show how to integrate these mask constraints into different kinds of existing background subtraction techniques and demonstrate our results on 2 specific algorithms: a pixel-based and a MRF-based algorithm. Any other algorithm of such type or a patch-based algorithm could be modified in such way as well.

3. MASK MODEL

Let I be a video frame, $M_{x,y}$ – mask of the foreground layer, $M_{x,y} = 1$ for foreground pixels, and $M_{x,y} = 0$ for background pixels.

Considering the described scenario and the shape of objects of interest (people), we propose to limit possible foreground masks to one of the 2 following mask models:

Model 1

- **Formal definition**: for every pixel column x, there is such $h(x)$, so that $M_{x,y} = 1$ for every $y \le h(x)$ and $M_{x,y} = 0$ for every $y > h(x)$.

So if any pixel is marked as foreground then all pixels under it must also be marked as foreground. An example of such mask is demonstrated on Fig 1 (c), 2 (c, g). As can be seen, it is rather adequate for examples, where a person is standing or walking frontally (doesn't bend). Most errors occur in cases of bending or outstretched arms.

Model 2

- **Formal definition**: for every pixel column x, there are such $h_1(x)$ and $h_2(x)$, $h_1(x) < h_2(x)$ so that $M_{x,y} = 1$ for every $y \in [h_1(x), h_2(x)]$ and $M_{x,y} = 0$ for $y < h_1(x)$ or $y > h_2(x)$.

Examples of this model are given on Fig. 1 (d), Fig. 2 (d, h). This model places fewer constraints on the foreground mask. As can be seen in Fig 2 (h), this mask is much closer to the ground truth foreground mask, than mask of model 1. The remaining errors occur when there are more, than 2 foreground/background transitions in a single pixel column – for example near a neck (from bottom to top: shoulder-background near a neck – face – background again).

3.1 Model validation

To validate the proposed mask models we conducted the following experiments. We transformed ground truth segmentations of the 100 manually annotated frames from 20 different video sequences, so that they conformed to our 2 models. For the Model 1 we just marked every pixel under any foreground pixel as foreground, and for the Model 2 we found highest and lowest foreground pixels for every column, and marked every pixel between them as foreground.

Then we computed a number of misclassified pixels. In Table 1 we show these results compared to results, obtained by common pixel-based and MRF-based approaches.

As can be seen, number of misclassified pixels for the transformed masks is much less than for the algorithms results, which leaves us a hope, that by constraining the algorithms to produce only masks of these types we could achieve a better segmentation.

	Misclassified pixels, %
Ground truth – Model 1	2.4
Ground truth – Model 2	1.0
Pixel-based algorithm	4.3
MRF-based algorithm	4.0

Table 1. Number of misclassified pixels comparison between background subtraction results and the proposed mask models. See text for the details.

4. PROPOSED ALGORITHM

In this section we show, how to constrain pixel-based and MRF-based algorithms to produce only the proposed kind of masks.

4.1 Pixel-based algorithm

Let $p(I_{x,y} \mid M_{x,y}))$ be a probability of pixel (x,y) to have color $I_{x,y}$, being labeled $M_{x,y}$.

Common pixel-based algorithms mark a pixel as foreground if $p(I(x,y) \mid 1) > p(I(x,y) \mid 0)$ and as background otherwise.

Now, let's consider the case of mask Model 1. In this case for every x we need to find $h(x)$ – such that all pixels above $h(x)$ belong to a background, and all pixels under $h(x)$ belong to a foreground. So we need to maximize the following likelihood:

$p(h(x)=Y)=p(I_1,...,I_{Y-1},I_Y,...,I_H \mid 0,...,0,1,...,1)$, where we have skipped index x, H - is a frame height, lower y-indices correspond to higher pixels. As $I_1,...,I_H$ are conditionally independent of mask labels M,

$$p(h(x)=Y)=\prod_{y<Y} p(I_y \mid 0)\cdot\prod_{y\geq Y} p(I_y \mid 1)\cdot$$

Instead of probability itself we can maximize its logarithm: $\ln p(h(x)=Y)=\sum_{y<Y}\ln p(I_y \mid 0)+\sum_{y\geq Y}\ln p(I_y \mid 1)$

For fast computation we just need to precompute following integral projection images:

$F_{x,Y}=\sum_{y\geq Y}\ln p(I_{x,y} \mid 1)$ and $B_{x,Y}=\sum_{y<Y}\ln p(I_{x,y} \mid 0)$

Now $\arg\max(\ln p(h(x)=Y))=\arg\max(F(x,Y)+B(x,Y))$ and we need only to find a maximum value in each column of a precomputed image.

In case of mask Model 2, for every x we need to find 2 values $h_1(x)$ and $h_2(x)$. Now the likelilhood is $p(h_1(x)=Y_1,h_2(x)=Y_2)=\prod_{y<Y_1} p(I_y \mid 0)\cdot\prod_{Y_1\leq y\leq Y_2} p(I_y \mid 1)\cdot\prod_{y>Y_2} p(I_y \mid 0)$

Loglikelihood is

$$\ln p(h_1(x)=Y_1,h_2(x)=Y_2)=B(x,Y_1)+(F(x,Y_1)-F(x,Y_2+1))+$$
$$(B(x,H+1)-B(Y_2+1))$$

We can optimize it either globally by searching through all possible pairs (Y_1,Y_2) or by searching for the best Y_1 at first (getting mask Model 1), fixing it, and searching for the best Y_2.

As can be seen from the algorithm, this modification adds very little computational overhead. Another advantage is no additional parameters.

4.2 Constraining the algorithm, based on MRF or CRF

These algorithms minimize an energy of the form $E(M)=\sum_p\phi(I \mid M_p)+\sum_{\{p,q\}\in N}\phi(I \mid M_p,M_q)$,

where $\phi(I \mid M_p)$ and $\phi(I \mid M_p,M_q)$ are some unary and pair-wise potentials. Minimization of this energy is equivalent to maximizing a maximum aposteori probability. This energy is usually globally minimized by means of binary graph cut.

To constrain this algorithm to produce only masks of Model 1, we perform the following:

- Add 2 artificial hidden rows of pixels (variables in the random field) to the image – one at the top of it, the other at the bottom.
- Modify unary potentials for the added rows: $\phi(I_p \mid 1)=Inf$ for the top row and $\phi(I_p \mid 0)=Inf$ for the bottom row.
- Vertical pair-wise potentials between the added rows and previous border rows is set equal to the nearest pair-wise vertical potential in the image.
- Add to all vertical pair-wise potentials some 'big value':
  $\phi'(I_p \mid M_p,M_q)=\phi(I_p \mid M_p,M_q)+BigValue$

Such unary potentials for the top and bottom rows 'oblige' the algorithm to make at least 1 cut in every image column, and modified vertical pair-wise potentials lead to maximum 1 cut in every image column. $BigValue$ must be larger than sum of unary potentials for any column. So the artificial bottom row is marked as foreground, top row is marked as background, and for every column there is a single label transition – and the mask is conformed with mask Model 1.

In case of Model 2, we perform operations described above, so that the mask becomes of type 1. Then we fix the top border of the foreground mask, and perform the same operation with foreground-background mask labels switch. However, this doesn't give the global minimum.

5. RESULTS AND COMPARISON

We have tested our modifications on 2 specific algorithms:

- A pixel-based algorithm, where a pixel background color probability is modeled as a single 3d Gaussian. Foreground color model is uniform.

  We have also experimented with different uniform foreground color probability values (so mathematically they are not probabilities, because they do not sum up to 1), which is equivalent to comparing background color probability to a threshold.
- MRF-based algorithm. Unary potentials are computed from the same pixel color probabilities, as in the pixel-based

algorithm. Pair-wise potential is standard for segmentations with GraphCut and uses background attenuation from [5].

For the experiments we have collected 20 video sequences, shot by a camera, placed at 1.3m – 2m above the floor level. Video resolution varies between 320x240 and 640x480, but we downsampled all video sequences to 320x240. Video recording was conducted in 3 different rooms with different illumination.

To be able to evaluate background subtraction results we have manually segmented 100 random frames into foreground and background.

In Table 2 we show the results of the experiments. For all algorithms (original and modified) we used the same parameters. It can be seen, that the proposed mask constraints improve segmentation accuracy for both algorithms. Both types of constraints decrease the algorithm precision, but increase the recall.

From Table 2 we can note, that on our data MRF-based algorithm made rather modest segmentation accuracy improvement, compared to a simple pixel-based approach. But modified versions of both algorithms showed almost similar results.

By changing foreground probability (i.e. threshold) we can make precision/recall curves for the pixel-based algorithm and its modifications (for MRF-based algorithm a dependency on this parameter is more complex). These curves are demonstrated on Fig. 3. We can see, that mask model 2 shows the best precision/recall ratio.

Algorithm	Misclassified pixels, %	Precision, %	Recall %
Pixel-based	4.3	96.4	80.5
Pixel-based, Mask Model 1	3.1	88.9	93.4
Pixel-based, Mask Model 2	2.9	94.7	88.6
MRF-based	4.0	96.8	81.4
MRF-based, Mask Model 1	3.0	90.2	92.8
MRF-based, Mask Model 2	3.4	95.4	84.3

Table 2. Mask accuracy comparison between original and the modified algorithms.

6. CONCLUSION

In this paper, we have considered a special case of video surveillance scenario, where a camera is attached to a wall on 1-2m height and records people walking by or coming up to the camera. We proposed 2 special foreground mask models for this scenario and showed how to integrate new mask constraints into typical background subtraction algorithms. The proposed modifications have no parameters and add little computational overhead. Experiments, conducted on our own video sequences, demonstrated segmentation accuracy improvement of the modified algorithms.

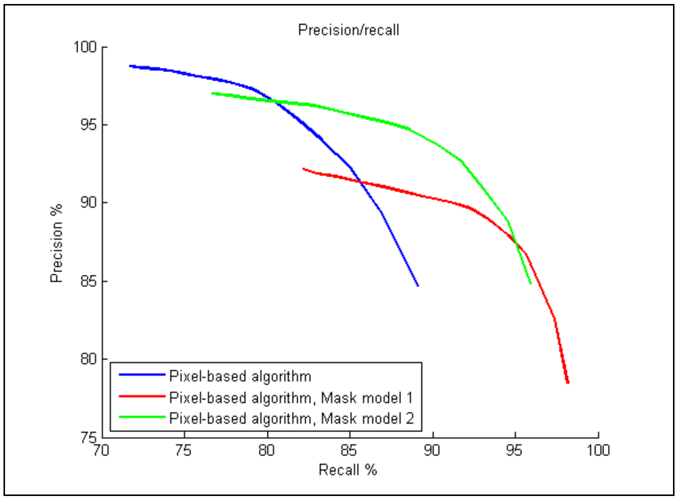

Figure 3. Precision/recall curves for the original pixel-based background subtraction algorithm and its 2 modifications. See text for the details.

7. ACKNOWLEDGEMENTS

This work was supported by Russian Foundation for Basic Research, project N. 09-01-92474-MNKS_a, Presidential Grant for the Support of Leading Scientific Schools N. NSh-8129.2010.9

8. REFERENCES

[1] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland, Pfinder: Real time tracking of the human body. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19 (7), pp. 780–785, 1997

[2] C. Stauffer and W. E. L. Grimson, Adaptive background mixture models for real-time tracking. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 246–252, 1999

[3] A. Elgammal, D. Harwood, and L. Davis, Non-Parametric Model for Background Subtraction, In *Proc. Sixth European Conf. Computer Vision (ECCV),* 2, pp. 751-767, 2000

[4] M. Heikkila, M. Pietikainen, J. Heikkila, A texture-based method for detecting moving objects, In *British Machine Vision Conference (BMVC)*, 2004

[5] J. Sun, W. Zhang, X. Tang, and H. Y. Shum, Background cut, In *Proc. Europ. Conf. on Computer Vision (ECCV)*, pp. 628-641, 2006

[6] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov, Bilayer segmentation of live video, In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 53–60, 2006

[7] P. Yin, A. Criminisi, J. Winn, and I. Essa, Tree-based classifiers for bilayer video segmentation, In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR),* 2007

[8] Y. Boykov and M. Pi. Jolly, Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *Proceedings of ICCV*, pp. 105–112, 2001