PER-PIXEL BACKGROUND ESTIMATION IN VIDEO MICROSCOPY USING FRAME GROUPING AND WAVELET BASED IMAGE FUSION

Chi Cui

Department of Electrical and Computer Engineering, University of Maryland, College Park
email: cch2000@gmail.com

ABSTRACT

Per-pixel background estimation is very important for accurate tracking of biological structures in video microscopy. Mixture modeling of the background is hard due to less priors and heavy halo effects around the target biological structures in each frame. Since halos could have similar brightness and motion pattern to the foreground sometimes they are not identified as in the background by these methods. In this paper we proposed a new method based on incremental frame grouping and iterate image fusion to compute the per-pixel background in video microscopy. Experiments on simulated microscopy videos and real sequence data set show that our method could calculate the background with a quite high precision. At the same time it also outperforms previous methods in removing strong noise and heavy image blurring, the "halo", which are especially useful attributes for video microscopies.

1. INTRODUCTION

Per-pixel background estimation is a very important problem for accurate tracking of biological structures in video microscopy. Since we have less priors for the micro-structures than for the daily objects, such as pedestrians, air planes, cars etc, it is usually quite difficult to model either the foreground or the background properly. For video microscopy the background is usually a soup of different kinds of noise. Sometimes there are also artifacts caused by the defects on the surface of the capture device or the cover slip. Most of the background estimation in microscopy video done previously works on a mono-frame basis which doesn't take advantage of the temporal coherency information. The background in a single frame is either cleared out by filters or modeled using a mixture of Gaussian. Recent work [2][3][5]on the second way shows fine results on the background subtraction of life scenes, such as traffic and pedestrians. However it could not work equally well on video microscopies since these methods tend to fail in identifying the halos which are quite common in video microscopy as background. All these halos are bind to the boundaries of objects and they are quite bright, sometimes even brighter than parts of the cellular structures, and they are not static. Most of the time they will move with the surrounded object and thus become very likely to be thresholded as the foreground in mixture modeling. However, in order to accurately quantify some cellular structures that the biologists are currently interested in, such as F-actin removing them is a must. Newly published STLBP method[1] also shows promising results on dynamic background modeling which could deals with moving branches or waving flags in the scene. However it doesn't suit for video microscopy processing either since the frames often contain sub-pixel structures, such as microtubule filaments. This makes the STLBP which works on per pixel basis not so powerful. In this paper we proposed a method that could do the per-pixel background estimation based on the image fusion from a set of similar frames. Incoming frames are first grouped by some criterion and then the background within the group of frames is estimated by wavelet image fusion. Section 2 gives a detailed description of our algorithm. Some experimental results on simulated data and real microscopy video sequences using our method are presented in section 3. The comparison between the original frames and background free results shows that our approach could obtain the background estimation which contains both halos and noise with a quite high precision. We draw a conclusion in the last section.

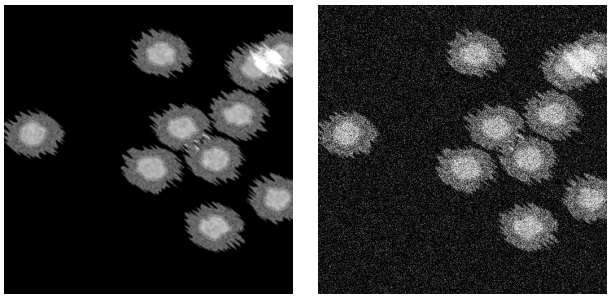

Figure 1: The figure on the left is a randomly chosen frame from the clean video (ground truth) and the one on the right is the same frame from the noise contaminated and blurred video for simulation.

2. PROPOSED METHOD

2.1 Frame Grouping

We did the background estimation on a frame group basis.The size of the frame group is not fixed however there is a fixed upper bound for computational efficiency. We kept a temporary data structure with a fixed length called frame grouping buffer.The incoming frames were first stored into the frame grouping buffer. When it was full, frames in the buffer were clustered into frame groups based on an optimization function to be introduced later. If the group assignment resulted in a singleton and it was an early frame this frame is either assigned to the previous group or the current larger group depending on its similarity to its neighboring frame in the corresponding group. Otherwise the group containing early frames were removed from the buffer and the remaining frames were pushed back to the bottom of the buffer to make space for new incoming frames. We employed Kingsburys Dual-Tree Complex Wavelet Transform

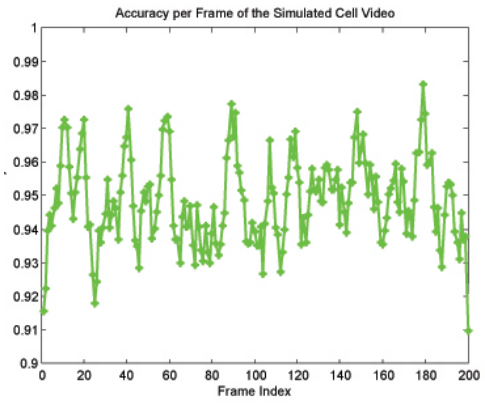

Figure 2: This figure shows the accuracy of background estimation on a frame basis of the whole simulation video sequence

(DT-CWT)[10] to compute a redundant image representation with six oriented complex detail subbands at each decomposition level for each frame. The advantages of this complex wavelet transform variant are its approximate shift-invariance, its directional selectivity and the very efficient implementation scheme by four parallel 2-D DWTs. All of these properties come at the very low cost of four-times redundancy in 2-D. The dissimilarity between consecutive frames was computed as the root mean square of the difference between wavelet coefficient magnitudes over all the levels shown in Eq.1

$$DS_{i,j} = \sum_{c=1}^{L}(\sum_{k=1}^{6}(RMS(I_{c,k}^{i}, I_{c,k}^{j}))) \quad (1)$$

where L denotes the composition level and k denotes the index of bandpass.$Ii_{c,k}$ denotes the component image of k^{th} bandpass at composition level c of frame i. In order not to make the frame group too small we only cut the frames in the frame grouping buffer into 2 pieces. We used $DS_{i,j}$ to denote the dissimilarity between frame i and frame j then the decision of frame intersection point could be turned into the following optimization problem shown in Eq.2

$$p_s = argmax_i(\omega DS_{i-1,i} + max_{j<i}((1-\omega)DS_{j,i})) \quad (2)$$

where ω was a weight to balance the consecutive frame dissimilarity and the maximum of all the previous dissimilarity.If p_s was computed as i, the intersection was between i-1 and i. The first part in p_s is trying to find the abrupt changes and the second part is a representation of accumulated difference from gradual changes. We are seeking to strike a balance between these factors by solving the optimization function above. We also kept a threshold for the optimized intersection dissimilarity function Eq.3

$$f_i = \omega DS_{i-1,i} + max_{j<i}((1-\omega)DS_{j,i}) \quad (3)$$

if f_{opt} was below this threshold all the frames in the frame grouping buffer was assigned to the same group otherwise they were divided up as the intersection point location. Since we need to compute the pairwise dissimilarity the selection

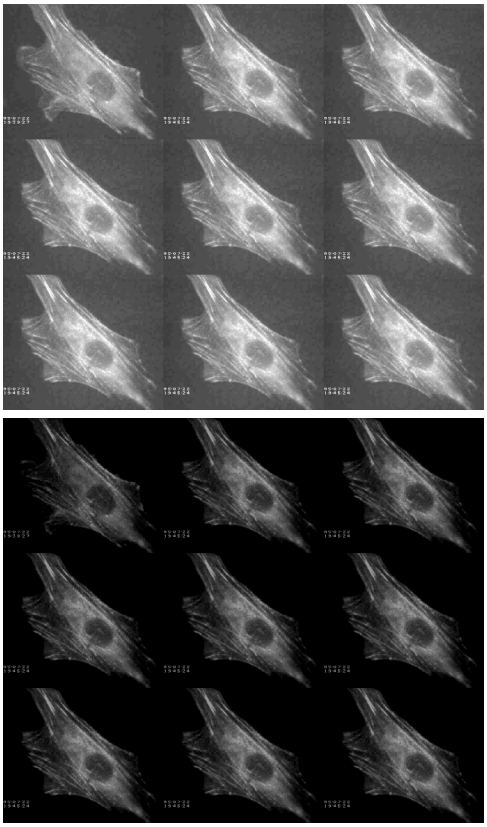

Figure 3: This figure shows an example of our experiment results on mono-color real microscopy videos. The video shown here is downloaded from [7]. 9 consecutive randomly chosen frames were put in one montage. The image on the top is the original frame collection and the one that follows is the post processed frame collection.

of frame grouping buffer size is crucial to both the computational efficiency and the accuracy of our algorithm. In all of our experiments that follows this depth was set to five based on trails and errors.

2.2 Group Based Background Estimation

All the frames in a group used the same background estimation. The background was constructed by initially setting it to pure black then literately doing image fusion on the current background and one frame in the group.The background is updated with the image fusion result each time. The iteration stopped until it had gone though all the frames in the group and the final image fusion result was the estimated background. We did the image fusion based on wavelet decomposition. The procedure mainly followed what is introduced in [6].

- The two images are respectively decomposed into sub-images using forward wavelet transform, which have the same resolutions at the same level and different resolution among different levels and
- Information fusion is performed based on both the high-frequency and low-frequency sub-images of decomposed images; and finally the result image is obtained using inverse wavelet transform.

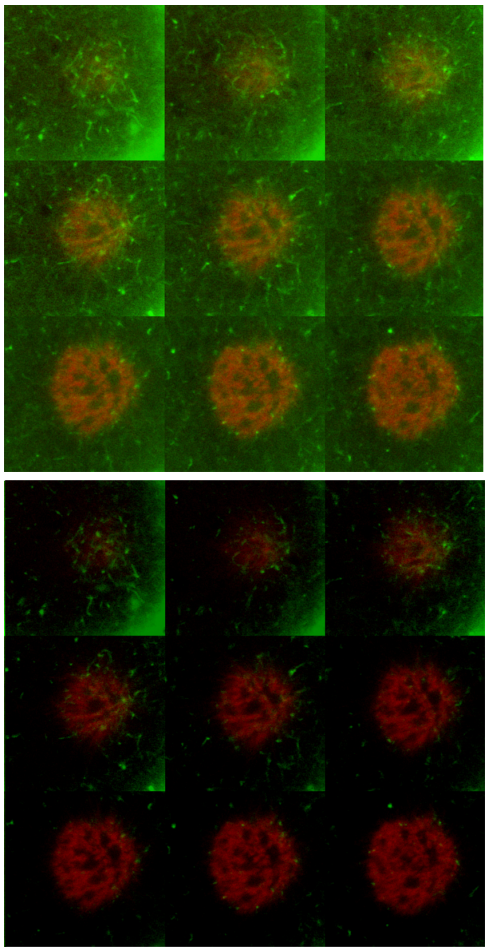

Figure 4: This figure shows an example of our experiment results on color real microscopy videos. The video shown here was downloaded from [8]. 9 consecutive randomly chosen frames were put in one montage. The image on the top is the original frame collection and the one that follows is the post processed frame collection.

The activity level measurement is pixel based. However we made some modification to the combining methods. Besides of choosing the max of the coefficients from the images to be fused, we added choosing the min into the combination strategy. Choosing the mean of coefficients from images to be fused was just a specified version of weighted average. Let $I_1(x,y)$ and $I_2(x,y)$ be the images to be fused, the decomposed low-frequency sub-images of $I_1(x,y)$ and $I_2(x,y)$ which contains the main content of the image at resolution s be represented as $I^l_{1,s}(x,y)$ and $I^l_{2,s}(x,y)$ respectively. Similarly the decomposed high frequency sub-images which contains the details of the image at resolution s be represented as $I^h_{1,s}(x,y)$ and $I^h_{2,s}(x,y)$. Noticing that in 2D images the high frequency components might refer to more than 1 band, we just use the same annotation here out of simplicity consideration. The low frequency component of the fused image at pixel location p, namely $J^l_s(p)$ could be computed as shown in E.q.4

$$J^l_s(p) = [\frac{\sum_{q\in N_p} I^l_{1,s}(q)}{|N_p|} + \frac{\sum_{q\in N_p} I^l_{2,s}(q)}{|N_p|}]/2 \qquad (4)$$

Where N_p denotes the neighborhood of pixel location p. The high frequency component of the fused image at pixel location p, namely $J^h_s(p)$ could be computed as shown in E.q.5

$$J^h_s(p) = min(\frac{\sum_{q\in N_p} I^h_{1,s}(q)}{|N_p|}, \frac{\sum_{q\in N_p} I^h_{2,s}(q)}{|N_p|}) \qquad (5)$$

As seen in the Eq.4 and Eq.5 the low frequency component was approximated by the mean of the two images in a local neighborhood while the the high frequency components were approximated by the minimum of the two images in a local neighborhood. This was because low frequency part usually contained the key structures of the image so a mean was used to over come the gaussian blurring effects, which appear as halos. The high frequency part was more likely to contain noise and thus a minimum operator was used. To make the computation efficient and preserve the information of details we used haar wavelet and the decomposition level was set up to 10. Though Daubechies and Symlets also appear to be reasonable choices, experiments show that the simplest wavelet basis, Haar wavelet works the best in all the cases. This corresponds to the conclusion in [11] for edge detection. Even though the background was generated on a small group frames, say less than or equal to 5, it was still a small number compared to the number of frames in most of our experiments and thus the calculation was quite fast. The background needed to be estimated for each image channel separately. Therefore in the RGB cases, 3 backgrounds should be computed and subtracted from the corresponding color channel respectively.

3. EXPERIMENT RESULTS AND ANALYSIS

3.1 Simulation

We generated a 200 frame simulated video of cell motion. The main strategy for generating the cellular structures followed the work [9] done by Lehmussola, A. et al. The clean frames were first contaminated with Poisson noise, then convoluted by a 5 by 5 gaussian kernel with the standard deviation of 0.5 to approximate the cellular blurring effect. We added the gaussian noise with 0 mean and 0.01 variance, assuming the pixel intensities were scaled to range between 0 and 1, to simulate the noise from capture devices. One frame of the resultant video was shown in Fig.1. We applied our method to estimate the per-pixel background in each frame. The accuracy of the background estimation is computed as in Eq.6

$$\tilde{A} = 1 - [\frac{\sum_i \frac{|I^f_i - \hat{I}^f_i|}{I^f_i}}{N} + \frac{\sum_i \frac{|I^b_i - \hat{I}^b_i|}{I^b_i}}{N}]/2 \qquad (6)$$

where N denotes the number of pixels in the image, I^f_i denotes the foreground intensity of pixel i and $\hat{I}^f_i$ denotes the estimated value. Similarly I^b_i denotes the background intensity of pixel i and $\hat{I}^b_i$ denotes the estimated value. The first part in E.q.6 represents the accuracy for the estimation of the foreground and the second term shows the accuracy

for the estimation of the background. The accuracy measure based on Eq.6 of the processed video based on this setting was shown in Fig.2. Most of the accuracies were around 95 percent which was a quite good result for per-pixel based background estimation.

3.2 Real Video

We also applied our per-pixel background estimation algorithm to videos from the real microscopy data, both RGB and mono-color. The gray scale one had 59 frames with a resolution at 280 by 236 and was downloaded from the company website [7]of the work [6]done by Yu-li Wang. It showed the ATP deplection that causes an immediately retraction of all the lamellipodia in which the actin filament bundles (stress fibers) collapse into tight aggregates over 1-2 hours. The RGB color video was one of the supplement media resources for the paper[8] done by J. Azoury et al. It contained 39 frames and the resolution was 550 by 534. It showed the confocal sections of the F-Actin Cage around the Microtubule Spindle. Z stack of a spindle from a control oocyte expressing Utr-GFP (green) and Map7-RFP (red) observed by live confocal microscopy at BD + 8 hr. The spindle axis is perpendicular to the plane of observation. Z step, 1 m. We picked two videos containing different kinds of biological structures. This was to test whether our method had structure dependencies. The validation was performed manually by the visual inspection of domain experts. Fig.3 and Fig.4 showed a montage for processed and unprocessed consecutive frame images from these two videos. The starting points were randomly chosen. The two images on the left were the 9 frames result of the mono-color video before and after subtracting the estimated background using our method. And the two on the right in the same figure showed the 9 frames of the RGB video before and after subtracting the estimated background. In the unprocessed images of the mono-color frame sequence, the halos could be seen on the boundaries of the cells and the fibers traversing the inside the cells were seriously blurred. All these artifacts were removed in the background free image montage on the left. Also the heavy background noise in the green channel of the RGB frame sequence was successfully cleaned which resulted in a much clearer view of the red nuclei. Strong noise, halos, which were hard to be removed as background by Mixture Modeling, now could be easily dealt with.

4. CONCLUSION

We presented a per-pixel background estimation method by first clustering the incoming frames into frame groups using the root mean squared difference between the magnitude of DT-CWT coefficients of two consecutive frames then iteratively doing image fusion between the constructed background and each frame in the group. The background was updated by the image fusion result after each iteration. Simulation on computer generated video sequence and real experiments on microscopy videos showed the algorithm could estimate the background in a quite high precision and demonstrated the special advantages of our method in removing halos and strong noise in the background which were often difficult for the mixture modeling related approaches.

REFERENCES

[1] A. Shimada, D. Arita, and R. Taniguchi,"Dynamic control of adaptive mixture-of-gaussians background model,"in *Video and Signal Based Surveillance, 2006. AVSS 06. IEEE International Conference on*, pp. 5C5, Nov. 2006.

[2] Dar-Shyang Lee, J.J. Hull, and B. Erol, "A bayesian framework for gaussian mixture background modeling,"in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 3, pp. IIIC973C6 vol.2, Sept. 2003.

[3] Zoran Zivkovic and Ferdinand van der Heijden, Efficient adaptive density estimation per image pixel for the task of background subtraction, *Pattern Recogn. Lett.*, vol. 27, no. 7, pp. 773 – 780, 2006.

[4] Shengping Zhang, Hongxun Yao, and Shaohui Liu, "Dynamic background modeling and subtraction using spatio-temporal local binary patterns,"in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pp. 1556 – 1559, Oct. 2008.

[5] Nick Kingsbury, A dual-tree complex wavelet transform with improved orthogonality and symmetry properties, 1998, pp. 319 – 322.

[6] Zhong Zhang; Blum, R.S., A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application,*Proceedings of the IEEE*, vol.87, no.8, pp.1315-1326, Aug 1999.

[7] A. Lehmussola, P. Ruusuvuori, J. Selinummi, T. Rajala, and O. Yli-Harja, "Synthetic images of high-throughput microscopy for validation of image analysis methods,"in *Proceedings of the IEEE*, vol. 96, no. 8, pp. 1348 – 1360, Aug. 2008.

[8] Y. l.Wang, , http://ylwang.umassmed.edu/video/index.htm.

[9] Y. l. Wang, Effects of atp depletion on actin organization, *Exp. Cell Res.*, vol. 167, pp. 16C28, 1986.

[10] Virginie Georget Pascale Rassinier Benjamin Leader Jessica Azoury, Karen W. Lee and Marie-Hlne Verlhac, Spindle positioning in mouse oocytes relies on a dynamic meshwork of actin filaments, *Current Biology*, vol. 18, no. 19, pp. 1514C1519, October 2008.

[11] S. Rajeev, R.E. Vasquez, R. Singh, Comparison of Daubechies, Coiflet, and Symlet for edge detection, *Visual Information Processing VI*, vol. 3074, SPIE, 1997, pp. 151C159.