Counting people in the crowd using a generic head detector

Venkatesh Bala Subburaman Adrien Descamps Cyril Carincotte
Image Department, Multitel absl, 7000 Mons, Belgium
{balasubburaman, descamps, carincotte}@multitel.be

Abstract

Crowd counting and density estimation is still one of the important task in video surveillance. Usually a regression based method is used to estimate the number of people from a sequence of images. In this paper we investigate to estimate the count of people in a crowded scene. We detect the head region since this is the most visible part of the body in a crowded scene. The head detector is based on state-of-art cascade of boosted integral features. To prune the search region we propose a novel interest point detector based on gradient orientation feature to locate regions similar to the top of head region from gray level images. Two different background subtraction methods are evaluated to further reduce the search region. We evaluate our approach on PETS 2012 and Turin metro station databases. Experiments on these databases show good performance of our method for crowd counting.

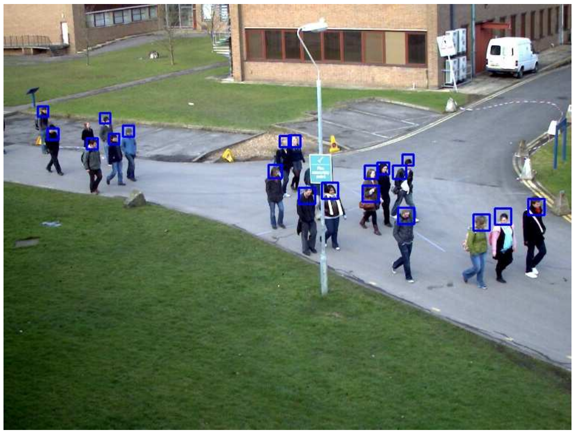


Figure 1: Example detection result from our approach. The blue boxes represents the detected heads from the image.

1 Introduction

Crowd counting has been an important component in video surveillance systems. For example, different levels of attention could be based on crowd of different density [1]. There are different approaches for counting the number of people from a crowded scene. One of the approaches is to directly detect humans from an image or videos. There has been a vast amount of work done in the area of human detection [2], and it usually performs well in scenes that are not very crowded. The other approach is indirect method which is based on a regression function to map the features from the region to the person count [3, 4, 5, 6, 7]. The two approaches have certain advantages depending on the scene considered. Detection based approach provides the exact location of the person, but might require sufficient resolution for reliable detection and would be challenging to detect multiple people in crowded/occluded situation. Whereas regression based approaches could be trained to work on low image resolution and crowded situations, but cannot provide the exact location of person, and tends to be scene specific.

In this paper we investigate detection based approach for crowd counting task on PETS 2012 and Turin metro station databases. An example from PETS database is shown in Figure 1. Given a crowded situation, it is natural for humans to count the number of people based on the visible part of the body which is mainly the head region. This motivates us to face the problem of crowd counting based on detecting the head rather than the whole human body. Detection based approach also has the advantage that it can be applied on a single image rather than a video. To reduce the search space for detecting the head, we propose a novel interest point detector based on gradient orientation feature that detects the top of head region. We also experiment methods to further reduce the search region with two different background subtraction techniques.

The rest of the paper is organized as follows. In the next section we briefly review the related work. In section 3, we give an overview of the proposed approach, followed by a description of our novel interest point and the head detector. In section 4 we evaluate our approach on benchmark databases, and in section 5, we conclude the paper.

2 Previous Work

There are different approaches that have been proposed for estimating the people count from an image and videos. Most of the approaches use background subtraction to segment humans from the background. In [3, 4] background subtraction is used to obtain foreground pedestrian pixels which are mapped to the number of people in the scene. The issue is as the crowd increases the estimation will deviate from true density due to occlusion and it is not easy to figure out just by foreground pixels if it belongs to the same or different person. In [8, 9], the authors mapped some feature statistics extracted from the blob to the count of people. In [5], a feature response belonging to the object is given a weight such that the sum of feature weights extracted from a full object sums to one. Thus even if partial object is visible their method will give partial count to the object.

In [10], maximally stable extremal regions (MSER) is used as a region extraction prototype for crowd feature description. In [11], corner points based on Kanade-Lucas-Tomasi are extracted which are clustered to detect human like structures after removing any background features by using foreground mask.

In [12], top of head region is detected after background subtraction to segment roughly the human by an ellipsoid. To reduce the dependence on an accurate foreground contour, which may be easily corrupted by noise, Rittscher et al. [13] extracted some additional feature points from the contour. A variant of Expectation Maximization (EM) is used to group these features into some human-sized rectangles. The authors have shown that their method work well even under low resolutions. However, if the background model depends on pixel value average over time, as used in their paper, obtaining good foreground might be difficult when the people in the scene are stationary.

In [6, 7], SURF features are extracted and only those which are in motion are considered. A regression function is used to map the number of SURF features to the people count.

Head detection using skeleton graph for people counting is proposed in [14]. The skeleton graph is extracted from the foreground mask obtained using background subtraction. In [15], a SVM classifier is trained to detect contours of head region, and then a perspective transform technique is used to estimate the crowd size more accurately. Our work is more closely related to method in [15] and follows a detection based approach to count people from an image.

3 Overview of our approach

Our approach mainly relies on a head detector to count people from an image. To detect the heads from the image we first find interest points using gradient information from the gray scale image which approximately locates top of the head region to reduce the search space. Our approach of locating top of the head is different from [13], and [12], where they find the top of head from foreground segmented image, which might not work well in highly crowded regions. The interest points on the image are masked using a foreground region obtained using background subtraction techniques such as Vibes[1] [16] and Idiap[2] [17]. A sub-window is then placed around the interest points, based on perspective calibration information, and it is classified as head or non-head region using a classifier. Multiple nearby detections are finally merged to obtain final number of detections.

3.1 Interest points based on gradient orientation

The proposed interest point detector is based on quantized gradient orientation which is simple and also fast to compute. Given an image I the gradient in x (g_x) and y (g_y) direction is computed by:

$$g_x(i,j) = I(i,j-1) - I(i,j+1) \tag{1}$$

$$g_y(i,j) = I(i-1,j) - I(i+1,j) \tag{2}$$

where i and j represents image index. The magnitude (M) and the orientation (O) are given by:

$$M(i,j) = \sqrt{g_x(i,j)^2 + g_y(i,j)^2} \tag{3}$$

$$O(i,j) = tan^{-1}\frac{g_y(i,j)}{g_x(i,j)} \tag{4}$$

Next, a binary image (B) based on the gradient orientation O is created, which is given by:

$$B(i,j) = \begin{cases} 1 & \text{if } O(i,j) \in [\pi_l, \pi_u] \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where π_l and π_u represents the lower and upper gradient orientation values. Given the binary image B, connected component labelling [18] is applied to find different regions. The centroid of each region forms our interest point.

Two parameters control the number of interest point generated by our approach. One is the threshold on the gradient magnitude M, and the other is the threshold on the region area in the binary image B.

3.2 Background subtraction

Given an image sequence, background subtraction can be applied to reduce the search region and thus speeding up processing time for detecting human head in an image.

[1] http://www2.ulg.ac.be/telecom/research/vibe/

[2] http://www.vanaheim-project.eu/downloads/resources.php

We consider Vibes and Idiap background subtraction techniques that are publicly available. Vibes uses a very simple approach and it is computationally less expensive than Idiap background subtraction method, whereas Idiap background subtraction takes advantages of local texture features represented by local binary patterns (LBP) and photometric invariant color measurements in RGB color space. The parameter setting for background subtraction methods can result in varied foreground segmentation results. Note that we use the background subtraction to reduce the search space, and our approach does not require precise segmentation of the foreground as required by other indirect approaches, thus easing the parameter setting for background subtraction. Finally, a simple morphological operation was applied to fill the gaps after foreground segmentation.

3.3 Head detector

The head detector is based on adaboost classifier, combined with a soft cascade that optimize the computation time by discarding rapidly easy negative sample. The features used are based on integral channel features [19]. For an input image, a set of channels are computed. In our case we compute the LUV channels, the gradient intensity channel and six oriented gradient channels. The features are generated by the sum of channels pixels on a random region of interest. A large set of random features are generated and adaboost selects the most relevant ones. The computation of the feature is made very efficient by the use of integral images. For more detailed description of the algorithm, see [19].

Training a head detector requires a high number of positive and negative training samples. Positive samples are usually collected by manual annotation of bounding box around heads. Similar to [5], we use annotations that are just a point on the center of the head or the face region. The negative data is cropped from the background from the same images that are used for obtaining the positive samples. About 2000 positive and negative samples were extracted this way from annotations on one training view. All the image patches are resized to 31x31 pixels.

Note that the training data for the head were extracted only from various indoor surveillance data (different from the data processed in this paper) and no data from PETS 2012 or Turin database were used in training for this study.

4 Experiments

We evaluate our approach on PETS 2012 S1 database (view 1) and on Turin[3] metro CCTV footage for people counting task. The Turin database contains 6 different views, each consisting of 721 frames. The head center locations were manually annotated as a point for PETS S1 database and for Turin metro station database to obtain the ground-truth people count. Metro station data are usually challenging because some people are stationary for a long period of time waiting for a train, and therefore we decided not to use any background subtraction method for this database. We only used interest point and head detector to estimate the count from the Turin database.

Experimental setup: The head detector consists of 1000 weak-classifier which are selected sequentially using adaboost algorithm. To speed up the training process only 5000 random features are considered at each stage of boosting, and Multi-instance pruning is finally applied to obtain an efficient cascade detector. For the interest point detector, the gradient magnitude threshold (m_t) is set to 3 and regions having minimum number of pixels (r_t) is set to 2. The low values are chosen since the head size in PETS database were small. The values could be higher if the head size are larger and could drastically reduce the search region depending on the scene. We evaluate our approach with three different setting for the background subtraction. 1) Vibes: on-line background learning (color) 2) Idiap: the background is learnt off-line using the training data from view 1, and 3) without any background subtraction. The results are reported for the regions R0, R1 and R2 as described in PETS2012[4].

Performance measures: The performance is measured in terms of Mean Absolute Error (MAE) and the Mean Relative Error (MRE) defined as in [7]:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |G(i) - T(i)| \tag{6}$$

$$MRE = \frac{1}{N}\sum_{i=1}^{N} \frac{|G(i) - T(i)|)}{T(i)} \tag{7}$$

where N is the number of frames for the test sequence and $G(i)$ and $T(i)$ are the detected and the true number of persons in the i-th frame respectively.

In a metro station the crowd usually increases, at the departure or at the arrival of a train, and is usually low at other times. For evaluating the performance on Turin metro station database, we also report correlation value, which gives us a similarity measure of the crowd pattern evolution over time.

PETS 2012 crowd counting results: We provide the performance measure for the three different background setting

[3] http://www.vanaheim-project.eu

[4] http://www.pets2012.net

in Table 1. We compare the results with previously best published results in Table 2 with indirect method of counting [7]. Notice that the other approaches did not provide the results for 14-33 and 14-31. We can observe that our approach surprisingly has better performance for the three sequences 13-59, 14-06, and 14-17 (note that the ground truth we annotated is for every 7 frames).

Case	R0	R1	R2
Vibes Background subtraction			
S1.L1.13-57	5.95(30%)	1.9(19%)	2.5(32%)
S1.L1.13-59	2.08(11%)	1.86(18%)	0.86(11%)
S1.L2.14-06	2.4(12%)	1.4(10%)	1.7(11%)
S1.L2.14-17	2.2(10%)	1.89(10%)	0.79(9%)
S1.L3.14-31	7.0(31%)	2.8(34%)	2.5(20%)
S1.L3.14-33	16(52%)	13(44%)	1.0(13%)
Idiap Background subtraction			
S1.L1.13-57	6.17(35%)	2.27(21%)	2.95(36%)
S1.L1.13-59	2.08(14%)	1.64(18%)	1.02(16%)
S1.L2.14-06	3.01(13%)	2.35(14%)	1.42(10%)
S1.L2.14-17	2.9(11%)	2.21(15%)	1.22(11%)
S1.L3.14-31	9.03(36%)	4.51(42%)	2.78(21%)
S1.L3.14-33	10.04(35%)	7.12(29%)	1.13(14%)
Without Background subtraction			
S1.L1.13-57	4.8(25%)	2.4(32%)	3.4(42%)
S1.L1.13-59	4.7(35%)	2.6(47%)	1.1(21%)
S1.L2.14-06	5.0(60%)	3.1(74%)	3.3(23%)
S1.L2.14-17	7.8(35%)	2.88(56%)	1.4(13%)
S1.L3.14-31	5.6(24%)	2.8(37%)	3.6(63%)
S1.L3.14-33	2.9(30%)	8.02(45%)	1.27(26%)

Table 1: Counting Performance measures with different background settings

Case	Albiol [20]	Percanella [7]	Ours
S1.L1.13-57	2.8(12.6%)	**1.36(6.8%)**	5.95(30%)
S1.L1.13-59	3.8(24.9%)	2.55(16.3%)	**2.08(11%)**
S1.L2.14-06	5.14(26.1%)	5.40(20.8%)	**2.4(12%)**
S1.L2.14-17	2.64(14.0%)	2.81(15.1%)	**2.2(10%)**
S1.L3.14-31	-	-	7.0 (31%)
S1.L3.14-33	-	-	16(52%)

Table 2: Counting performance measures compared with two other reported results

We only show the smoothed count plot for sequences S1.L1.13-57, S1.L2.14-17, and S1.L3.14-33 in Figures 2-5. Sequences S1.L1.13-57 and S1.L2.14-31 contains people walking away from the camera, while sequences S1.L2.14-06, S1.L1.13-59, and S1.L3.14-17 contains people walking toward the camera. We notice from Figure 2 that the estimated count is lower than the ground-truth count when compared with Figure 3, and a similar behaviour is observed for other sequences. The reason which we think is that the training data for the head detector contained people mainly facing towards the camera and not the back of the head. Figure 4 shows the estimated count results for Time-14-33. In this particular image sequence, people approach the center from different directions and stop for a while in the center. The count drops from frame number 150, due to the fact that the foreground estimation gets worse as the people are stationary for a while at the same location. The count increases again when people start to move from frame number 300. With Idiap background subtraction where the background is learnt off-line we can see that there is improvement in the estimated count as shown in Figure 5 when the people are stationary for a time period (see Figure 6).

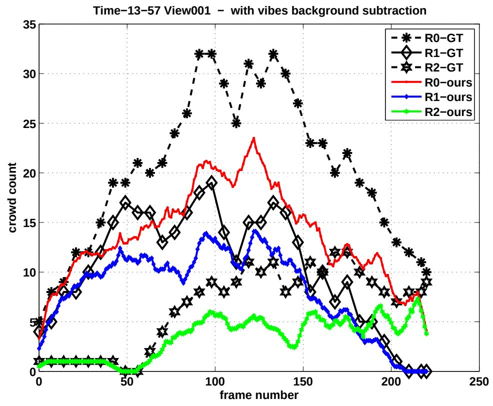


Figure 2: Plot of counting results for S1.L1.13-57 view 001 with Vibes background subtraction method.

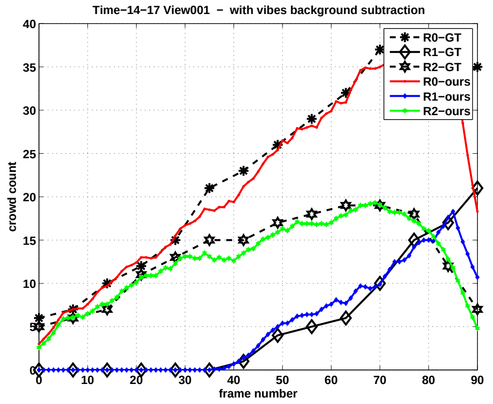


Figure 3: Plot of counting results for S1.L3.14-17 view 001 with Vibes background subtraction method.

Turin metro platform counting results: Finally, our algorithm was evaluated on platform views from Turin

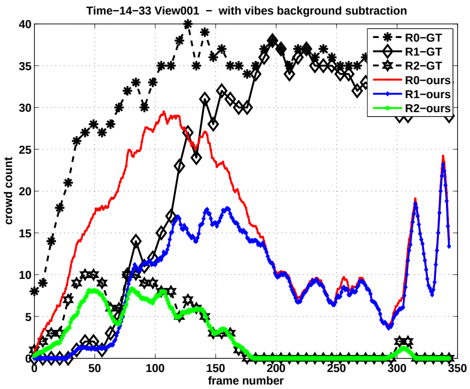


Figure 4: Plot of counting results for S1.L3.14-33 view 001 with Vibes background subtraction method.

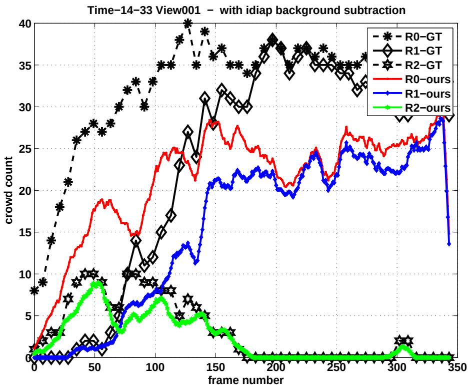
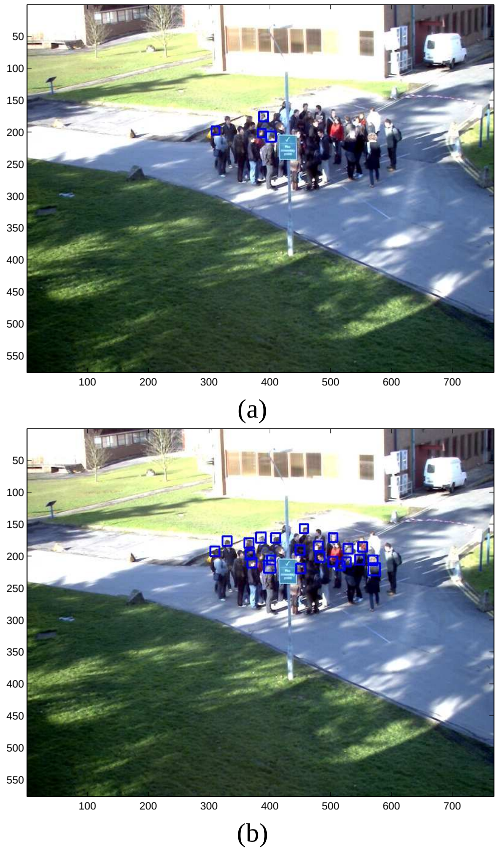


Figure 5: Plot of counting results for S1.L3.14-33 view 001 with Idiap background subtraction method.

database. It is worth to note that these views were not used in training. There are six different views of the two metro platforms, with close and far views. Examples from Turin database are shown in Figure 7.

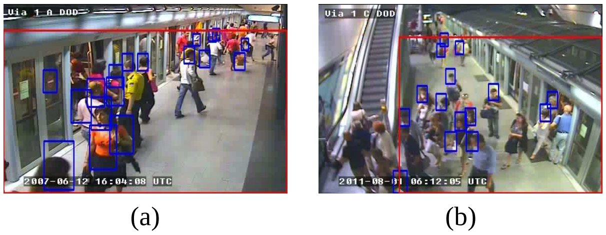
Figure 8, shows the smoothed version of the crowd evolution over time for these views (Via1A and Via1C), and performance for the six views are reported in Table 3. Though there are false and miss detections at frame level, the average estimated count follows the ground-truth count nicely. Table 3 shows the performance for Turin database and we summarize the observations next. Results for far views (Via1C and Via2A) are slightly affected by the higher difficulty to detect heads in low resolutions. Some other specific problems are observed : the view Via2B presents some stable false alarms in background, and view Via2C present a lower detection rate due to capture condition. Despite these limitations, the estimated counts follow quite accurately the periodic evolution observed in ground truth in all views.



(a)

(b)

Figure 6: Detection example for frame number 248 with a) Vibes background subtraction and b) Idiap background subtraction on 14-33 time slot.



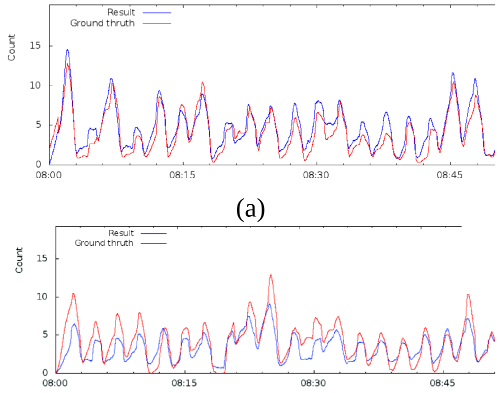
(a)

(b)

Figure 7: Detection example from Turin database for Via1A (a) and Via1C (b). The blue boxes represents the result of head detection, red box is the detection area.

5 Conclusions

We presented a method for counting crowd in images using head detection with interest points based on gradient orientation. Experiments on PETS and Turin databases show the potential for such an approach in different conditions of people moving in the scene, and the same head detector was tested on both the databases. There was also no fine tuning of the background subtraction parameters to obtain



(a)

(b)

Figure 8: Result of counting people in Turin Via1A (a) and Via1C (b). We have only shown for a portion of time slot.

View	Mean number of persons	Number of frames	Adaboost-Integral Channel	
			MAE	Correlation
Via1A	3.8	721	0.53	0.92
Via1B	3.6	721	1.20	0.95
Via1C	3.6	721	1.08	0.87
Via2A	2.4	721	0.88	0.68
Via2B	2.3	721	3.22	0.86
Via2C	3.0	721	1.36	0.85

Table 3: Performance of our counting algorithms for different views

the best results in case of PETS and no background subtraction was used in Turin database, which makes our approach more ideal to estimate the crowd from a single frame. Nevertheless, there is a lot of scope for improvement in our approach. In our future work we would like to reduce the false detections, and incorporate additional methods to reason out occlusions in a crowded scene.

Acknowledgement The research leading to these results has received funding from the European Communitys Seventh Framework Programme FP7/2007-2013 - Challenge 2- Cognitive Systems, Interaction, Robotics - under grant agreement n248907-VANAHEIM.

References

[1] A. B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, “Crowd analysis: a survey,” *Machine Vision and Applications*, pp. 345–357, 2008.

[2] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art,” *PAMI*, vol. 34, no. 4, pp. 743–761, 2012.

[3] Y. J., V. S., and A. Davies, “Image processing techniques for crowd density estimation using a reference image,” in *ACCV*, vol. 3, 1995, pp. 6–10.

[4] R. Ma, L. L., H. W., and T. Q., “On pixel count based crowd density estimation for visual surveillance,” in *IEEE Conf. Cybernet. Intell. Syst.*, vol. 1, 2004.

[5] V. Lempitsky and A. Zisserman, “Learning to count objects in images,” *NIPS*, 2010.

[6] D. Conte, P. Foggia, G. Percannella, F. Tufano, and M. Vento, “An effective method for counting people in video-surveillance applications.” in *VISAPP*, 2011, pp. 67–74.

[7] D. Conte, P. Foggia, G. Percannella, and M. Vento, “A method based on the indirect approach for counting people in crowded scenes,” in *AVSS*, 2010, pp. 111–118.

[8] Q. Huihuan, W. Xinyu, and X. Yangsheng, “Intelligent surveillance systems,” 2011, book : Springer.

[9] A. B. Chan and N. Vasconcelos, “Privacy preserving crowd monitoring: Counting people without people models or tracking,” *CVPR*, 2008.

[10] S. Hang, Y. Hua, and Z. Shibao, “The large-scale crowd density estimation based on effective region,” in *ACCV*, vol. 3, 2010, pp. 302–313.

[11] Y.-L. Hou, S. Member, G. K. H. Pang, and S. Member, “People counting and human detection in a challenging situation,” *IEEE Transactions on Systems Man and Cybernetics Part A Systems and Humans*, vol. 41, no. 1, pp. 24–33, 2011.

[12] Z. Tao and N. Ram, “Tracking multiple humans in complex situations,” *PAMI*, pp. 1208–1221, 2004.

[13] J. Rittscher, P. H. Tu, and N. Krahnstoever, “Simultaneous estimation of segmentation and shape,” in *CVPR*, 2005, pp. 486–493.

[14] K. E. Aziz, D. Merad, B. Fertil, and N. Thome, “Pedestrian head detection and tracking using skeleton graph for people counting in crowded environments,” in *2th IAPR Conference on Machine Vision Applications (MVA2011)*, 2011.

[15] S.-F. Lin, J.-Y. Chen, and H.-X. Chao, “Estimation of number of people in crowded scenes using perspective transformation,” *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, pp. 645–654, 2001.

[16] O. Barnich and M. Van Droogenbroeck, “ViBe: A universal background subtraction algorithm for video sequences,” *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, June 2011.

[17] J. Yao and J. M. Odobez, “Multi-layer background subtraction based on color and texture,” in *CVPR*, 2007.

[18] L. He, Y. Chao, K. Suzuki, and K. Wu, “Fast connected-component labeling,” *Pattern Recognition*, vol. 42, no. 9, pp. 1977–1987, 2009.

[19] P. Dollar, Z. Tu, P. Perona, and S. Belongie, “Integral Channel Features.” BMVC, 2009.

[20] A. Albiol, M. J. Silla, and J. M. Mosi, “video anaysis using corner motion statistics.” in *Workshop on PETS*, 2009, pp. 31–38.