AUTOMATIC CROWD ANALYSIS FROM VERY HIGH RESOLUTION SATELLITE IMAGES

Beril Sirmacek, Peter Reinartz

German Aerospace Center (DLR), Remote Sensing Technology Institute
PO Box 1116, 82230, Wessling, Germany
(Beril.Sirmacek, Peter.Reinartz)@dlr.de

Commission VII

KEY WORDS: Very high resolution satellite images, Crowd detection, DEM, Local features, Probability theory, Shadow extraction, Road extraction

ABSTRACT:

Recently automatic detection of people crowds from images became a very important research field, since it can provide crucial information especially for police departments and crisis management teams. Due to the importance of the topic, many researchers tried to solve this problem using street cameras. However, these cameras cannot be used to monitor very large outdoor public events. In order to bring a solution to the problem, herein we propose a novel approach to detect crowds automatically from remotely sensed images, and especially from very high resolution satellite images. To do so, we use a local feature based probabilistic framework. We extract local features from color components of the input image. In order to eliminate redundant local features coming from other objects in given scene, we apply a feature selection method. For feature selection purposes, we benefit from three different type of information; digital elevation model (DEM) of the region which is automatically generated using stereo satellite images, possible street segment which is obtained by segmentation, and shadow information. After eliminating redundant local features, remaining features are used to detect individual persons. Those local feature coordinates are also assumed as observations of the probability density function (pdf) of the crowds to be estimated. Using an adaptive kernel density estimation method, we estimate the corresponding pdf which gives us information about dense crowd and people locations. We test our algorithm using Worldview-2 satellite images over Cairo and Munich cities. Besides, we also provide test results on airborne images for comparison of the detection accuracy. Our experimental results indicate the possible usage of the proposed approach in real-life mass events.

1 INTRODUCTION

Recently automatic detection of people and crowds from images gained high importance, since it can provide very crucial information to police departments and crisis management teams. Especially, detection of very dense crowds might help to prevent possible accidents or unpleasant conditions to appear. Due to their limited coverage of area, street or indoor cameras are not sufficient for monitoring big events. In addition to that, it is not always possible to find close-range cameras in every place where the event occurs.

Due to the importance of the topic, many researchers tried to monitor behaviors of people using street, or indoor cameras which are also known as close-range cameras. However, most of the previous studies aimed to detect boundaries of large groups, and to extract information about them. The early studies in this field were developed from closed-circuit television images (Davies et al., 1995), (Regazzoni and Tesei, 1994), (Regazzoni and Tesei, 1996). Unfortunately, these cameras can only monitor a few square meters in indoor regions, and it is not possible to adapt those algorithms to street or airborne cameras since the human face and body contours will not appear as clearly as in close-range indoor camera images due to the resolution and scale differences. In order to be able to monitor bigger events researchers tried to develop algorithms which can work on outdoor camera images or video streams. Arandjelovic (Arandjelovic, Sep. 2008) developed a local interest point extraction based crowd detection method to classify single terrestrial images as crowd and non-crowd regions. They observed that dense crowds produce a high number of interest points. Therefore, they used density of SIFT features for classification. After generating crowd and non-crowd training sets, they used SVM based classification to detect crowds. They obtained scale invariant and good results in terrestrial images. Unfortunately, these images do not enable monitoring large events, and different crowd samples should be detected before hand to train the classifier. Ge and Collins (Ge and Collins, 2009) proposed a Bayesian marked point process to detect and count people in single images. They used football match images, and also street camera images for testing their algorithm. It requires clear detection of body boundaries, which is not possible in airborne images. In another study, Ge and Collins (Ge and Collins, 2010) used multiple close-range images which are taken at the same time from different viewing angles. They used three-dimensional heights of the objects to detect people on streets. Unfortunately, it is not always possible to obtain these multi-view close-range images for the street where an event occurs. Chao et al. (Lin et al., Nov. 2001) wanted to obtain quantitative measures about crowds using single images. They used Haar wavelet transform to detect head-like contours, then using SVM they classified detected contours as head or non-head regions. They provided quantitative measures about number of people in crowd and sizes of crowd. Although results are promising, this method requires clear detection of human head contours and a training of the classifier. Unfortunately, street cameras also have a limited coverage area to monitor large outdoor events. In addition to that, in most of the cases, it is not possible to obtain close-range street images or video streams in the place where an event occurs. Therefore, in order to behaviors of large groups of people in very big outdoor events, the best way is to use airborne images which began to give more information to researchers with the development of sensor technology. Since most of the previous approaches in this field needed clear detection of face or body features, curves, or boundaries to detect people and crowd boundaries which is not possible in airborne images, new approaches are needed to ex-

tract information from these images. Hinz et al. (Hinz, 2009) registered airborne image sequences to estimate density and motion of people in crowded regions. For this purpose, first a training background segment is selected manually to classify image as foreground and background pixels. They used the ratio of background pixels and foreground pixels in a neighborhood to plot density map. By observing change of the density map in the sequence, they estimated motion of people. Unfortunately, their approach did not provide quantitative measures about crowds. In a following study (Burkert et al., Sep. 2010), they used previous approach to detect individuals. Positions of detected people are linked with graphs. They used these graphs for understanding behaviors of people.

In order to bring a fully automatic solution to the problem, we propose a novel framework to detect people from remotely sensed images. One of the best solutions to monitor large mass events is to use airborne sensors which can provide images with approximately 0.3 m. spatial resolution. In previous studies (Sirmacek and Reinartz, 2011a) and (Sirmacek and Reinartz, 2011b), we used airborne images to monitor mass events. In the first study (Sirmacek and Reinartz, 2011a), we proposed a novel method to detect very dense crowd regions based on local feature extraction. Besides, detecting dense crowds, we have also estimated number of people and people densities in crowd regions. In following study (Sirmacek and Reinartz, 2011a), by applying a background control, individual persons are also detected in airborne images. Moreover, in a given airborne image sequence, detected people are tracked using Kalman filtering approach. Although airborne images are useful to monitor large events, unfortunately sometimes flying over mass event might not be allowed, or it might be an expensive solution. Therefore, detecting and monitoring crowds from satellite images can provide crucial information to control large mass events. As the sensor technology is being developed, new satellites can provide images with higher spatial resolutions. With those new satellite sensors, it became possible to notice human crowds, and even individual persons in satellite images. Therefore, herein we propose a novel approach to detect crowds automatically from very high resolution satellite images. Although resolutions of satellite images are still not enough to see each person with sharp contours, we can still notice a slight change of intensity and color components at the place where a person exists. Therefore, the proposed algorithm is based on local features which are extracted from intensity and color bands of the satellite image. In order to eliminate redundant local features which are generated by the other objects or texture on building rooftops, we apply a feature selection method which consists of three steps as; street classification approach, eliminating high objects on streets using shadow information, and using digital elevation model (DEM) of the region which is automatically generated using stereo satellite images to eliminate buildings. After applying feature selection, using selected local features as observations, we generate a probability density function (pdf). Obtained pdf helps us to detect crowded regions, and also some of the individual people automatically. We test our algorithm using Worldview-2 satellite images which are taken over Cairo and Munich cities. Our experimental results indicate the possible usage of the proposed approach in real-life mass events and to provide a rough estimation of the location and size of crowds from satellite data. Next, we introduce steps of the approach in detail.

2 LOCAL FEATURE EXTRACTION

In order to illustrate the algorithm steps, we pick $Munich_1$ image from our dataset. In Fig. 1.(a), we represent original $Munich_1$ panchromatic WorldView-2 satellite test image, and in Fig. 1.(b), we represent a subpart of this image in order to give information about real resolution. As can be seen here, satellite image resolutions do not enable to see each single person with sharp details. On the contrary, each person is represented with two or three mixed pixels, and sometimes additionally two or three mixed shadow pixels. All those pixels coming from a human appearance make a change of intensity components at the place where the person exists which can be detected with a suitable feature extraction method. Therefore, our crowd and people detection method depends on local features extracted from input image.

(a)

(b)

Figure 1: (a) $Munich_1$ test image from our Worldview-2 satellite image dataset, (b) Real resolution of a small region in $Munich_1$ test image.

For local feature extraction, we use features from accelerated segment test (FAST). FAST feature extraction method is especially developed for corner detection purposes by Rosten et al. (Rosten et al., Nov. 2010), however it also gives high responses on small regions which are significantly different than surrounding pixels. The method depends on wedge-model-style corner detection and machine learning techniques. For each feature candidate pixel, its 16 neighbors are checked. If there exist nine contiguous pixels passing a set of pixels, the candidate pixel is labeled as a feature location. In FAST method, these tests are done using machine learning techniques to speed up the operation. For detailed explanation of FAST feature extraction method please see (Rosten et al., Nov. 2010).

We assume (x_i, y_i) $i \in [1, 2, ..., K_i]$ as FAST local features which are extracted from input image. Here, K_i indicates the maximum number of features extracted from panchromatic band of the input image. We represent locations of detected local features for $Munich_1$ test image in Fig. 2.(b). As can be seen in this image, we have extracted local features on street at places where each individual person exits. Unfortunately, many redundant features are also detected generally on building rooftops, and corners. For detection of people and crowds, first of all local features coming from other objects should be eliminated. For this purpose, we apply a feature selection method that we represent in

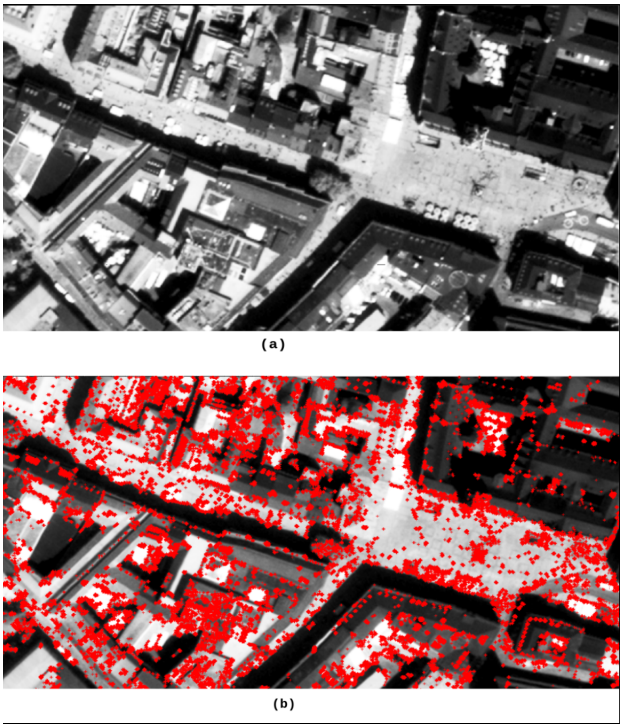


Figure 2: (a) Original $Munich_1$ test image, (b) FAST feature locations which are extracted from $Munich_1$ test image.

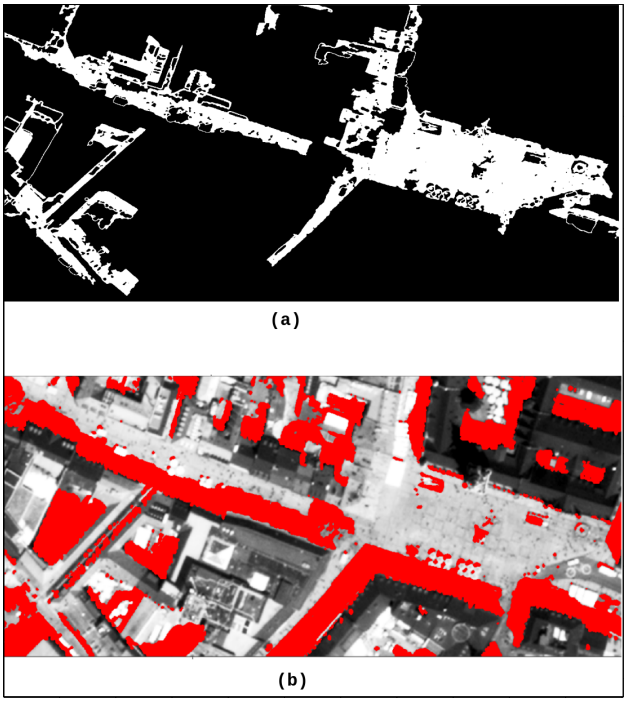


Figure 3: (a) Road-like pixels which are segmented from $Munich_1$ test image, (b) Automatically extracted shadow pixels from $Munich_1$ test image.

the next section in detail.

3 FEATURE SELECTION

For eliminating redundant features coming from building rooftop textures or corners of other objects in the scene, we use three masks as follows. The first mask ($B_1(x, y)$) is obtained by street segmentation using a training street patch which is selected by user. The second mask ($B_2(x, y)$), is generated using the shadow information, in order to remove high objects which appear on the detected street network. Finally, the third mask ($B_3(x, y)$) is obtained using height information obtained from DEM.

For street segmentation, we first choose a 20×20 pixel size training patch ($t(x, y)$) from input image. We benefit from normalized cross correlation to extract possible road segment. Normalized cross correlation between the training patch and the input image is computed using following equation.

$$\gamma(u,v) = \frac{\sum_{x,y}[g(x,y) - \overline{g}_{u,v}][t(x-u,y-v) - \overline{t}]}{\{\sum_{x,y}[g(x,y) - \overline{g}_{u,v}]^2 \sum_{x,y}[t(x-u,y-v) - \overline{t}]^2\}^{0.5}} \quad (1)$$

Here $\overline{t}$ represents the mean of intensity values in the template patch, and $\overline{g}_{u,v}$ represents the mean of the input image intensity values which are under the template image in correlation operation. At the normalized cross correlation result $\gamma(u, v)$, we obtain the road segment pixels as highlighted due to the high similarity to the training patch. By applying Otsu's automatic thresholding algorithm (Otsu, 2009) to the normalized cross correlation result, we obtain the road-like segments as in Fig. 3.(a). This binary image is assumed as the first mask ($B_1(x, y)$) which is going to be used for feature selection.

Although estimated street segment helps us for feature selection, still we cannot eliminate features coming from high objects on street such as street lamps, statues, small kiosks, etc. Unfortunately, those small objects also do not appear in DEM of the region, and they cannot be eliminated using height information coming from DEM. In order to eliminate features coming from these objects, in this step we try to detect them using shadow information. For shadow extraction, we use local image histograms. For each 100×100 pixel size window of the input image, the first local minimum in grayscale histogram is assumed as a threshold value to apply local thresholding to the image. After applying our automatic local thresholding method, we obtain a binary shadow map. In Fig. 3.(b), we represent detected shadow pixels on original image.

After detecting shadow pixels, we use the sun illumination angle to generate our high object mask. For labeling high objects, each shadow pixel should be shifted into opposite side of illumination direction. Assuming that (x_s, y_s) is an array of shadow pixel coordinates which are represented in Fig. 3.(b). New positions of shadow pixels $((\widehat{x}_s, \widehat{y}_s))$ are computed as $\overline{x}_s = x_s + l\sin(-\theta)$, and $\overline{y}_s = y_s + l\cos(-\theta)$. Here θ is the opposite direction of the illumination angle which is given by user, and l is the amount of shift in θ direction as pixel value. For better accuracy l should be chosen as the width of the shadow in illumination direction. However, in order to decrease computation time and complexity, we assume l equal to the length of the minor axis of an ellipse which fits shadow shape. After shifting shadow pixels, we generate our second mask $B_2(x, y)$ binary mask where $B_2(x, y) = 1$ for $((\widehat{x}_s, \widehat{y}_s))$. In Fig. 4, we illustrate shadow pixel shifting operation.

In order to obtain the last mask $B_3(x, y)$, we use DEM of the corresponding region which is generated from stereo Ikonos images using the DEM generation method of dAngelo et al. (dAngelo et al., 2009). We obtained $B_3(x, y)$ binary mask by applying local thresholding to DEM. We provide original DEM corresponding to $Munich_1$ image, and obtained binary mask in Fig. 5.(a), and (b) respectively. As can be seen, building rooftop regions are eliminated, however other low regions like park areas, parking

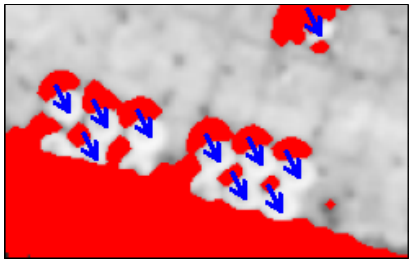


Figure 4: Illustration of shadow pixel shifting operation.

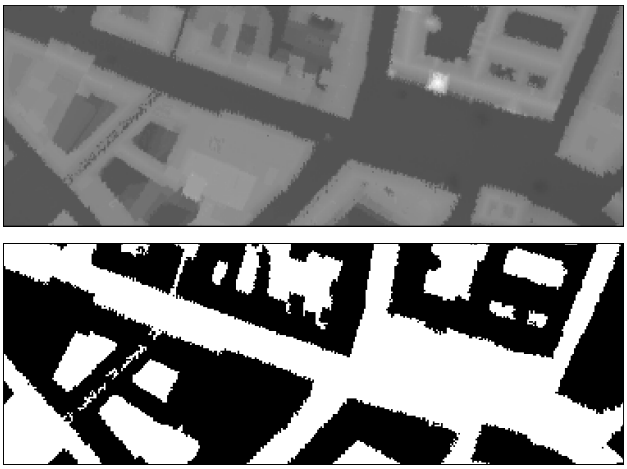


Figure 5: (a) Digital elevation model corresponding to $Munich_1$ test image which is generated using stereo WorldView-2 satellite images. (b) Low regions in $Munich_1$ image obtained by applying local thresholding to DEM.

lots with cars (or sea surface for some other test areas) cannot be eliminated with this mask. Therefore, we use information coming from three masks we generated. We assume our interest area as $S(x,y) = B_1(x,y) \wedge B_2(x,y)' \wedge B_3(x,y)$, where '$\wedge$' represents logical and operation for binary images.

We use detected $S(x,y)$ interest area for removing FAST features which are extracted from other objects. We eliminate a FAST feature which is at (x_i, y_i) coordinates, if $S(x_i, y_i) = 0$. Remaining FAST features behave as observations of the probability density function (pdf) of the people to be estimated. In the next step, we introduce an adaptive kernel density estimation method, to estimate corresponding pdf which will help us to detect dense people groups and also other people in sparse groups.

4 DETECTING INDIVIDUALS AND DENSE CROWDS

Since we have no pre-information about possible crowd locations in the image, we formulate the crowd detection method using a probabilistic framework. Assume that (x_i, y_i) is the ith FAST feature where $i \in [1, 2, ..., K_i]$. Each FAST feature indicates a local color change which might be a human to be detected. Therefore, we assume each FAST feature as an observation of a crowd pdf. For crowded regions, we assume that more local features should come together. Therefore knowing the pdf will lead to detection of crowds. For pdf estimation, we benefit from a kernel based density estimation method as Sirmacek and Unsalan represented for local feature based building detection (Sirmacek and Unsalan, 2010).

Silverman (Silverman, 1986) defined the kernel density estimator for a discrete and bivariate pdf as follows. The bivariate kernel function $[N(x,y)]$ should satisfy the conditions given below;

$$\sum_x \sum_y N(x,y) = 1 \tag{2}$$

$$N(x,y) \geq 0, \forall (x,y) \tag{3}$$

The pdf estimator with kernel $N(x,y)$ is defined by,

$$p(x,y) = \frac{1}{nh} \sum_{i=1}^{n} N(\frac{x - x_i}{h}, \frac{y - y_i}{h}) \tag{4}$$

where h is the width of window which is also called smoothing parameter. In this equation, (x_i, y_i) for $i = 1, 2, ..., n$ are observations from pdf that we want to estimate. We take $N(x,y)$ as a Gaussian symmetric pdf, which is used in most density estimation applications. Then, the estimated pdf is formed as below;

$$p(x,y) = \frac{1}{R} \sum_{i=1}^{K_i} \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma}) \tag{5}$$

where σ is the bandwidth of Gaussian kernel (also called smoothing parameter), and R is the normalizing constant to normalize $p_n(x,y)$ values between $[0, 1]$.

In kernel based density estimation the main problem is how to choose the bandwidth of Gaussian kernel for a given test image, since the estimated pdf directly depends on this value. For different resolution images, the pixel distance between two persons will change. That means, Gaussian kernels with different bandwidths will make these two persons connected to detect them as a group. Otherwise, there will be many separate peaks on pdf, however we will not be able to find large hills which indicate crowds. As a result, using a Gaussian kernel with fixed bandwidth will lead to poor estimates. Therefore, bandwidth of Gaussian kernel should be adapted for any given input image.

In probability theory, there are several methods to estimate the bandwidth of kernel functions for given observations. One well-known approach is using statistical classification. This method is based on computing the pdf using different bandwidth parameters and then comparing them. Unfortunately, in our field such a framework can be very time consuming for large input images. The other well-known approach is called balloon estimators. This method checks k-nearest neighborhoods of each observation point to understand the density in that area. If the density is high, bandwidth is reduced proportional to the detected density measure. This method is generally used for variable kernel density estimation, where a different kernel bandwidth is used for each observation point. However, in our study we need to compute one fixed kernel bandwidth to use at all observation points. To this end, we follow an approach which is slightly different from balloon estimators. First, we pick $K_i/2$ number of random observations (FAST feature locations) to reduce the computation time. For each observation location, we compute the distance to the nearest neighbor observation point. Then, the mean of all distances give us a number l. We assume that variance of Gaussian kernel (σ^2) should be equal or greater than l. In order to guarantee to intersect kernels of two close observations, we assume variance of Gaussian kernel as $5l$ in our study. Consequently, bandwidth of Gaussian kernel is estimated as $\sigma = \sqrt{5l}$. For a given sequence, that value is computed only one time over one

image. Then, the same σ value is used for all observations which are extracted from images of the same sequence. The introduced automatic kernel bandwidth estimation method, makes the algorithm robust to scale and resolution changes.

We use Otsu's automatic thresholding method on obtained pdf to detect regions having high probability values (Otsu, 2009). After thresholding our pdf function, in obtained binary image we eliminate regions with an area smaller than 1000 pixels since they cannot indicate large human crowds. The resulting binary image $B_c(x, y)$ holds dense crowd regions. Since our $Munich_1$ test image does not include very dense crowds, in Fig. 7 we illustrate an example dense crowd detection result on another Worldview-2 satellite test image which is taken over Cairo city when an outdoor event occurs.

After detecting dense crowds automatically, we focus on detecting individuals in sparse areas. Since they indicate local changes, we assume that detected local features can give information about individuals.

In most cases, shadows of people or small gaps between people also generate a feature. In order to decrease counting errors coming double counted people because of their shadows, we follow a different strategy to detect individuals. We use a binary mask $B_f(x, y)$ where (x_i, y_i) feature locations have value 1. Then, we dilate $B_f(x, y)$ using a disk shape structuring element with a radius of 2 to connect close feature locations. Finally, we apply connected component analysis to mask, and we assume mass center of each connected component as a detected person position. In this process, slight change of radius of structuring element does not make a significant change in true detected people number. However, an appreciable increase in radius can connect features coming from different persons which leads to underestimates.

5 EXPERIMENTS

To test the proposed algorithm, we use a Worldview-2 satellite image dataset which consists of four multitemporal panchromatic images taken over Munich city ($Munich_{1-4}$ images), and one panchromatic image taken over Cairo city ($Cairo_1$). Those panchromatic Worldview-2 satellite images have approximately half meter spatial resolution. We also test proposed algorithm on an airborne image (with 30 cm. spatial resolution) taken from the same region in over Munich city, in order to show robustness of the algorithm to resolution and sensor differences

In Fig. 6, we represent people detection results for $Munich_{1-4}$ images. For these four multitemporal images, true individual person detection performances are counted as $92,02\%$, $70,73\%$, $88,57\%$, and $89,19\%$ respectively. Besides, false alarm ratios are obtained as $14,49\%$, $40,34\%$, $24,29\%$, and $27,03\%$ respectively. In Fig. 7.(a), we present dense crowd detection and people detection results in Worldview-2 satellite image taken over Cairo city. Robust detection of dense crowd boundaries indicate usefulness of the proposed algorithm to monitor large mass events. Finally, in Fig. 7.(b), we represent people detection results on an airborne image which is taken in the same test area over Munich city. Obtained result proves robustness of the algorithm to scale and sensor differences of the input images.

6 CONCLUSION

In order to solve crowd detection and people detection, herein we introduced a novel approach to detect crowded areas automatically from very high resolution satellite images. Although resolutions of those images are not enough to see each person with sharp details, we can still notice a change of color components in the place where a person exists. Therefore, we developed an algorithm which is based on local feature extraction from input image. After eliminating local features coming from different objects or rooftop textures by applying a feature selection step, we generated a probability density function using Gaussian kernel functions with constant bandwidths. For deciding bandwidth of Gaussian kernel to be used, we used our adaptive bandwidth selection method. In this way, we obtained a robust algorithm which can cope with input images having different resolutions. By automatically thresholding obtained pdf function, dense crowds are robustly detected. After that, local features in sparse regions are analyzed to find other individuals. We have tested our algorithm on panchromatic Worldview-2 satellite image dataset, and also compared with an algorithm result obtained from an airborne image of the same test area. Our experimental results indicate possible usage of the algorithm in real-life events. We believe that, the proposed fully automatic algorithm will gain more importance in the near future with the increasing spatial resolutions of satellite sensors.

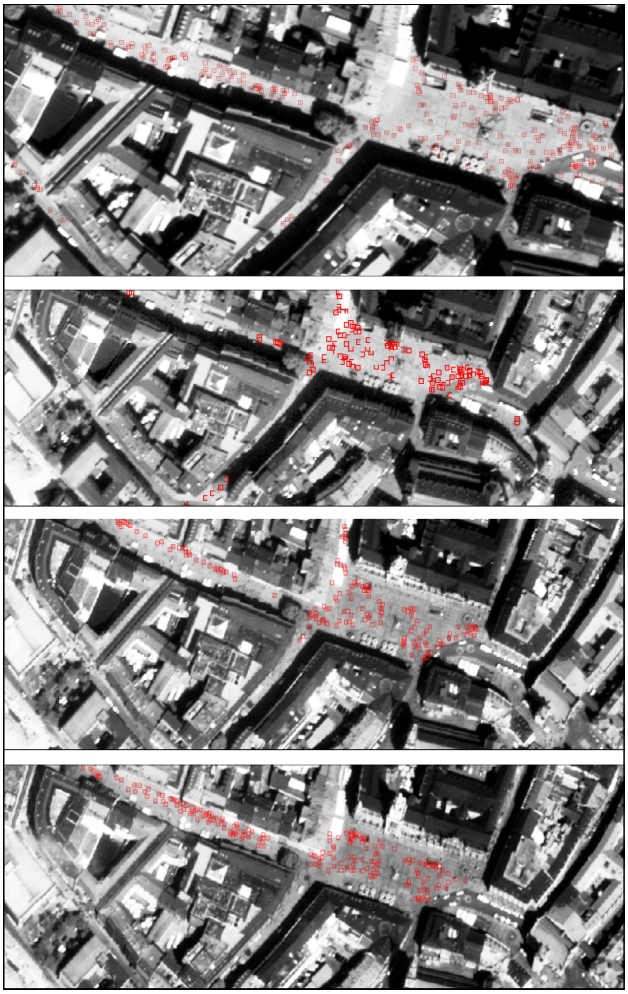
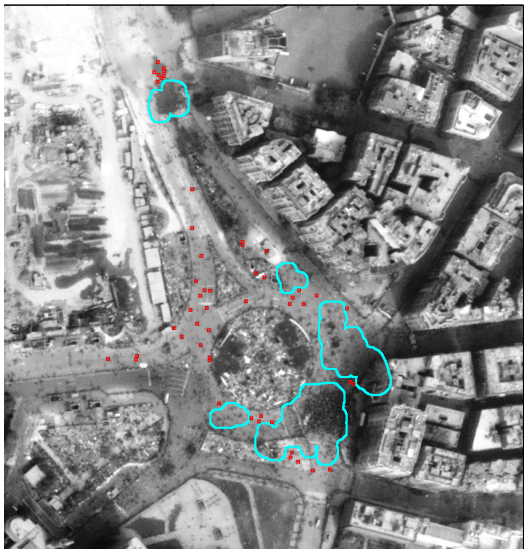


Figure 6: People detection results on $Munich_{1-4}$ Worldview-2 satellite images.

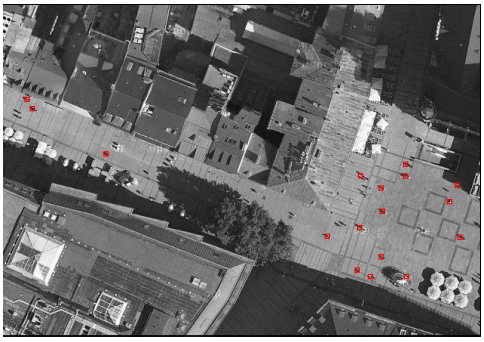
REFERENCES

Arandjelovic, O., Sep. 2008. Crowd detection from still images. British Machine Vision Conference (BMVC'08).

Burkert, F., Schmidt, F., Butenuth, M. and Hinz, S., Sep. 2010. People tracking and trajectory interpretation in aerial image sequences. International Archives of Photogrammetry, Remote Sensing and Spatial Infor-



(a)



(b)

Figure 7: (a) Dense crowd and people detection result on Worldview-2 satellite image taken over Cairo city, (b) People detection result on an airborne image which is taken at the same test area over Munich city.

mation Sciences (IAPRS), Commission III (Part A) XXXVIII, pp. 209–214.

dAngelo, P., Schwind, P., Krauss, T., Barner, F. and Reinartz, P., 2009. Automated dsm based georeferencing of cartosat-1 stereo scenes. In Proceedings of International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences.

Davies, A., Yin, J. and Velastin, S., 1995. Crowd monitoring using image processing. IEEE Electronic and Communications Engineering Journal 7 (1), pp. 37–47.

Ge, W. and Collins, R., 2009. Marked point process for crowd counting. IEEE Computer Vision and Pattern Recognition Conference (CVPR'09) pp. 2913–2920.

Ge, W. and Collins, R., 2010. Crowd detection with a multiview sampler. European Conference on Computer Vision (ECCV'10).

Hinz, S., 2009. Density and motion estimation of people in crowded environments based on aerial image sequences. ISPRS Hannover Workshop on High-Resolution Earth Imaging for Geospatial Information.

Lin, S., Chen, J. and Chao, H., Nov. 2001. Estimation of number of people in crowded scenes using perspective transformation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 31 (6), pp. 645–654.

Otsu, N., 2009. A threshold selection method from gray-level histograms. IEEE Transactions on System, Man, and Cybernetics 9 (1), pp. 62–66.

Regazzoni, C. and Tesei, A., 1994. Local density evaluation and tracking of multiple objects from complex image sequences. Proceedings of 20th International Conference on Industrial Electronics, Control and Instrumentation (IECON) 2, pp. 744–748.

Regazzoni, C. and Tesei, A., 1996. Distributed data fusion for real time crowding estimation. Signal Processing 53, pp. 47–63.

Rosten, E., Porter, R. and Drummond, T., Nov. 2010. Faster and better: A machine learning approach to corner detection. IEEE Transactions on Pattern Analysis and Machine Learning 32 (1), pp. 105–119.

Silverman, B., 1986. Density estimation for statistics and data analysis. 1st Edition.

Sirmacek, B. and Reinartz, P., 2011a. Automatic crowd analysis from airborne images. 5th International Conference on Recent Advances in Space Technologies RAST 2011, Istanbul, Turkey.

Sirmacek, B. and Reinartz, P., 2011b. Kalman filter based feature analysis for tracking people from airborne images. ISPRS Workshop High-Resolution Earth Imaging for Geospatial Information, Hannover, Germany.

Sirmacek, B. and Unsalan, C., 2010. A probabilistic framework to detect buildings in aerial and satellite images. IEEE Transactions on Geoscience and Remote Sensing.