A Scheme for the Detection and Tracking of People Tuned for Aerial Image Sequences*

Florian Schmidt and Stefan Hinz

Institute of Photogrammetry and Remote Sensing (IPF)
Karlsruhe Institute of Technology (KIT), 76128 Karlsruhe, Germany
{florian.schmidt,stefan.hinz}@kit.edu

Abstract. This paper addresses the problem of detecting and tracking a large number of individuals in aerial image sequences that have been taken from high altitude. We propose a method which can handle the numerous challenges that are associated with this task and demonstrate its quality on several test sequences. Moreover this paper contains several contributions to improve object detection and tracking in other domains, too. We show how to build an effective object detector in a flexible way which incorporates the shadow of an object and enhanced features for shape and color. Furthermore the performance of the detector is boosted by an improved way to collect background samples for the classifier training. At last we describe a tracking-by-detection method that can handle frequent misses and a very large number of similar objects.

Keywords: Aerial image sequences, object detection, classifier training, people tracking

1 Introduction

Aerial images sequences taken from high altitude allow to quickly overview wide areas and to analyze temporal changes. Automatic methods are needed to extract useful information in an appropriate manner out of this huge amount of digital data. This paper addresses this demand and proposes a method to detect and track a large number of people in aerial image sequences (Fig. 1). The results generated by our approach are especially useful in applications in which wide areas have to be monitored simultaneously as e.g. during mass events or for the evaluation of infrastructure.

1.1 Challenges and Contribution

The task of tracking individuals in aerial image sequences is very challenging and differs considerably from other domains due to the following reasons. A single person consists only of a few pixels in size and its appearance is influenced by

* This version of the paper has been created by the author. The original publication is available at www.springerlink.com.

Fig. 1. Part of a typical aerial image which we use in our analysis.

changing atmospheric conditions (Fig. 2). The number of individuals can vary from hundreds up to many thousands which all look very much alike. Aerial images include lots of person-like objects in particular in a complex urban environment. The sequences have a very low frame rate and are affected by the motion of the camera platform.

This paper explains how we handle these challenges. It contains the following key contributions:

1. We show how to build an effective detector for tiny objects by incorporating enhanced features in a flexible way which are tuned for certain characteristics as color, shape or shadow.
2. We present two methods to improve the way of collecting background samples for the classifier training.
3. A tracking-by-detection method is described which can handle frequent misses and which constructs reliable tracklets in a fast way for a very large number of similar objects.

1.2 Related Work

The use of image sequences taken from an airborne platform for surveillance tasks has been studied for many years. In most cases different methods for moving object detection are employed as in (Kumar et al., 2001), (Medioni et al., 2001), (Benedek et al., 2009) and (Reilly et al., 2010a). However these techniques are sensitive to the parallax effect and changing lighting conditions and do not work well for small or static objects. Appearance-based approaches overcome these problems since they work on single images. They have been successfully applied for car detection e.g. in (Yu et al., 2006), (Grabner et al., 2008) and (Leitloff et al., 2010).

Although there is an overwhelming amount of literature about the detection and tracking of people in terrestrial videos, only few publications exist which use

data of airborne platforms for this task. Xiao et al. (2008) presented a system which is based on moving object detection. Though they achieved reasonable results for vehicles, the algorithm failed for small and slow moving persons. In (Miller et al., 2008) Harris corner features are used to detect individuals and to avoid the motion dependency. Yet the poor results did not confirm their approach. Reilly et al. (2010b) achieved a good detection rate with a two stage procedure. Assuming that persons are upright shadow casting objects, they filter the image for candidate regions which fulfill this constraint. Afterwards the candidates are classified as human or clutter by using wavelet features. Although we utilize the shadow of persons as well, we integrate this information directly in an appearance-based framework. In Burkert et al. (2010) persons are detected by applying a sequence of segmentation algorithms. The resulting regions are tracked by calculating optical flow between consecutive images. It is questionable if these simple methods will yield robust results in more complex situations.

2 Overview of the Proposed Approach

We employ a tracking-by-detection framework to solve the problem of localizing and tracking individuals through a sequence of aerial images. The following sections describe in detail how we cope with the numerous challenges which arise in the domain of wide area surveillance. First we explain how we have designed an optimal detector for our object class (section 2.1) and how we have trained it in an efficient way (section 2.2). Afterwards we illustrate the tracking algorithm which links the previously generated detections (section 2.3).

2.1 Detector Design

One has to face particular problems when trying to detect individuals in images that have been taken in vertical direction from an altitude of, e.g. 1500 m. The size of a single person decreases to about 4 by 4 pixel at a ground sampling distance of 0.15 cm. Clouds and other atmospheric conditions can lower the signal-to-noise ratio even more, so that in some cases the shadow of a person is the only visible cue (Fig. 2).

Fig. 2. Example of a person with and without shadow at a common pixel size of 15 cm.

Most of the related work on object detection in aerial image sequences is based on algorithms for moving object detection. However these approaches do not work for static objects. They are furthermore unsuitable for very small and slow moving objects because of the moving camera and the improper alignment of the aerial images. For this reason we have decided to use an appearance-based approach instead in which features are extracted inside of a detection window and passed to a trained classifier to make a decision about the presence of an object. We use *Gentle AdaBoost* (Friedman et al., 2000) as classifier. This method has been successfully applied to detect very small objects e.g. by Leitloff et al. (2010) with cars in satellite images or by Smal et al. (2010) with spots in fluorescence microscopy images.

In general a sequence of complex preprocessing steps for geometric normalization and image alignment has to be applied to images taken from airborne platforms to perform further analysis. These tasks are not subject of this paper and have to be done in advance, c.f. (Kumar et al., 2001) or (Thomas et al., 2008). Our algorithm assumes orthorectified, georeferenced images of a fixed ground sampling distance as input data (Fig. 1). We execute nevertheless one object specific normalization which is explained in the following paragraph.

Incorporating the Shadow of a Person. An appearance-based detection approach works best if the feature values of one object class have a low variability in feature space. This can be achieved by choosing invariant features and by normalizing systematic influences. Since we use geometric normalized images from a nadir camera, single persons already have a similar dot-like shape (Fig. 2). The shadow casted by a person is an important additional cue for detection. Yet its appearance is highly variable and can only be modeled explicitly in the particular case when a single person stands isolated on a uniform ground like it has been done by Reilly et al. (2010b).

We propose to incorporate the potential object shadow directly in the appearance-based detection framework instead. This way we fuse information about a person and its potential shadow on the classifier level. The implicit model that is learned during the training of the classifier acts more flexible in ambiguous situation and shows better detection results.

Depending on the global position and the time of year and day the shadow of a person has different length and direction. Since we have georeferenced images, we use the available metadata to turn the images to correct for the different direction of the sun. Afterwards the shadows will always point in upward direction. This normalization step reduces the variance in object appearance. The benefit of geometric normalization exceeds clearly the minor information loss due to the necessary resampling.

We design the size of the detection window so that it covers the body of a single person and also a good part of its shadow (Tab. 1). In the training procedure we use samples of persons with shadow and without. The AdaBoost classifier has shown to be flexible enough to cope with this variation and combine all cues in the best way.

Object-specific Haar-like Features. Choosing the right features to describe the object of interest is a crucial step in appearance-based object detection. They should be invariant against noise and systematic variations and they should discriminate well between the object and background class. We have the additional requirement in our domain that the features have to be computable inside a very small detection window.

We use Haar-like features since they meet all stated requirements. They extract information about the shape of an object, they can be computed very fast at constant time in a window of at least 2 by 2 pixels and they are furthermore invariant to constant and linear changes in lighting (Viola and Jones, 2001).

The basic set of Haar-wavelet features introduced by Oren et al. (1997) has been extended to the more flexible Haar-like features (Viola and Jones, 2001), (Smal et al., 2010). This feature class can be generalized even more to better suit the shape of the object of interest (Fig. 3). Though these customized features can be assembled from the basic Haar-wavelet features, they lead to a faster detector since less features have to be extracted.

Fig. 3. A certain Haar-like feature is calculated by subtracting the sum of all pixel of the black from the white rectangle. The factor α is required to compensate for the differences in rectangle size. Five object-specific Haar-like features out of the 24 prototypes we have used to describe the specific shape of people in aerial images.

Rectangle Feature Class. Although shape provides the most dominating cues for object detection, we want to use color as complementary and supporting feature. Since our objects are very small, it is not feasible to extract a color histogram. Instead we could use the pixel values inside the detection window as feature values, however they would be very sensitive to noise.

For this reason we introduce the *rectangle feature class*. A single feature of this class is a value computed on a single channel in an arbitrarily shaped rectangle inside the detection window. We calculate mean and variance of all pixels inside a rectangle, as this can be done fast and analog to the Haar-like features. The mean value responds to color while the variance responds to homogeneity in object appearance. Since we want to utilize the rectangle features to represent the appearance of the object and not of its shadow, we use a smaller detection window as for the Haar-like features (Tab. 1).

Color Space and Window Size. The *rgb* color space is often not the best choice when processing color images because of the high correlation between

its channels. Instead we use the *i1i2i3* color space (Ohta et al., 1980) since it separates color and intensity information and is easy to compute. Table 1 displays an overview of our detector design. For every feature class we choose the optimal window size and color channel. By doing so, we get a detector which is particularly suitable for our object class. Furthermore the initial number of possible features is smaller which reduces the time for feature selection and training.

Table 1. Overview of the designed detector.

Feature class	Channel	Window size	Potential features
Haar-like	i1	9 x 15	9477
Rectangle	i1	9 x 9	3969
Rectangle	i2	9 x 9	3969
Rectangle	i3	9 x 9	3969

2.2 Training the Classifier

The training phase is essential in appearence-based object detection. Good samples for the object and background class are the base for good results. Object samples have to be collected manually by selecting individuals in training images. It is however far more difficult to gather appropriate background samples since this class is a priori boundless.

Usually the training samples should capture the entire distribution of feature values for both classes. However by using the binary AdaBoost classifier one just needs to find the right samples which define the boundary between object and background class in feature space. This can be done effectively by training the classifier in an iterative bootstrapping fashion (Sung and Poggio, 1998), (Grabner et al., 2008). We take all samples of the object class and a few manually marked background samples to start the training. We use a fixed number of stumps as weak learners. Afterwards we run the classifier on aerial images that do not contain any person. All detections in these images are therefor false positives and could be added to the set of background samples. Yet this would increase the number of negative samples very quickly to an unfeasible high number. So we add only very few randomly chosen false positives and repeat the training process until the number of false detections in images without persons decreases to a desired rate.

Confidence-aware Collection of Background Samples. One drawback of the described method is that other objects which look similar to the object of interest and which appear in the background images are added to the set of background samples. This general problem has a large effect in our application

domain since aerial images contain lots of person-like objects. Although the *Gentle AdaBoost* classifier (Friedman et al., 2000) is less sensitive to outliers in the training data, the best results are achieved if background and object class are not mixed.

We solve this problem by calculating a confidence measure for each detection. In general the AdaBoost detection score is unbounded and the decision for object or background class is based solely on the sign of the weighted sum of the answers of all weak learners. Yet the detection score can be converted to a more useful normalized confidence measure by dividing it by the sum of the weights of all weak learners (Grabner et al., 2008):

$$conf(\mathbf{x}) = \frac{\sum \alpha_n \cdot h_n^{weak}(\mathbf{x})}{\sum \alpha_n} \tag{1}$$

The confidence measure of each false positive detection can now be exploited to prevent a mixing of the training samples of object and background class. Instead of choosing false detections randomly, we prefer those with lowest confidence above 0. This strategy concentrates on ambiguous samples close to the decision border and reduces considerably the chance of mixing both classes during the iterative semi-automatic recording of background training samples.

Background Samples in Object Proximity. The detector would generate a lot of false detection if the associated classifier would have been trained exclusively with background samples from images without any objects. These false positives would occur especially in the close surrounding of the objects of interest, mainly in their shadow.

The reason for this behavior is that the classifier has not seen this area during training. We solve this problem by using also aerial images where the position of all existing persons has been marked. We collect background samples as described previously but do only consider all false detections which have a minimum distance to any marked person.

2.3 Detection and Tracking

We run the trained detector over a predefined region of interest inside an aerial image. The result is a confidence map where pixel values close to -1 indicate background and values close to $+1$ object positions. The confidence score has been calculated independently for every pixel. So we estimate the continuous two-dimensional confidence distribution with a Gaussian kernel. Afterwards we apply a non-maxima suppression to get a finite set of potential object positions. Only those candidate positions whose confidence score exceeds a fixed detection threshold are finally passed on to the tracking algorithm.

Tracking-by-detection with Frequent Misses. In our application domain we have to deal with frequent misses due to the fact that changing atmospherical

conditions can lower the signal-to-noise ratio to the point where single persons are hardly visible anymore. In addition, people often walk or stand in groups so that they merge visually to an undefined blob. Our detector does not work in these situations since it is trained to locate only isolated persons.

The tracking-by-detection method would fail if the percentage of misses is to high. Therefore we lower the detection threshold to reduce the miss rate to a minimum at the cost of more false alarms. The hard decision between background and object class is postponed in this way from the detection to the tracking stage where more information is available.

Tracking a Large Number of Objects in Clutter. Tracking in aerial images requires an algorithm that can handle lots of objects at the same time. This demand increases even more as we allow a lot of false detections during the detection stage. Further challenges arise because of a low frame rate of, e.g. 2 Hz and the fact the objects of interest have similar appearance. The true motion of individuals is furthermore affected by the inaccuracy in image alignment in a magnitude of few pixels.

We adapt an iterative Bayesian tracking approach similar to the one used by Betke et al. (2007) to track a large number of flying bats. A certain person is described using the following basic states:

1. position (x, y, confidence)
2. motion (direction, velocity)
3. color (r, g, b)

Information about position and color is provided by the detection algorithm. Additionally we calculate optical flow between consecutive images to extract motion information at the location of every observation. As most people move only a few pixels between two images, optical flow is well suited to get a good estimate for object motion.

A fast to compute near constant velocity motion model is used to predict object states for every new image. These predictions are linked afterwards with new observations on the base of state affinity. We apply the efficient gating strategy of Collins and Uhlmann (1992) to limit the number of possible associations and reduce the complexity of the data association problem. Objects with no assigned measurement are marked as *lost objects* and measurements without objects as *new objects*.

We analyze all connections and establish an affinity matrix for each cluster of connected components. The matrix elements represent affinity scores, which are calculated in a similar fashion as in (Wu and Nevatia, 2007). The affinity matrix serves as input for the the association module. We are looking only for one to one associations between objects and observations and prohibit split and merge situations.

There are several methods to solve this task. Some search for a global optimum by maximizing the overall association affinity while others make use of greedy, heuristic methods. We use the *direct link* method of Huang et al. (2008).

It does not provide a global optimum for the assignment problem but it generates reliable associations in a simple and fast way. A measurement is assigned to an object only if the affinity exceeds a certain minimum and only if the affinity to all other objects and measurements is considerable lower.

If an object is assigned successfully to an observation, its states are updated with a fast α-β-filter. Lost objects with no association are not tracked any further. This restriction ensures the non-overlapping constraint and prevents wrong associations in our domain with high object density and low frame rate.

In the final step of each iteration we calculate a track confidence for each object by averaging over the detection confidence of every position. If this score is below a certain threshold, the entire track is marked as clutter and discarded from further processing.

3 Experiments and Results

We have several aerial image sequences available to evaluate the proposed approach. They have all been taken with a camera facing in vertical direction from an altitude of about 1500 m resulting in a ground sampling distance of about 20 cm up to 15 cm. They show different big events with lots of people in an urban environment (Fig. 1). All sequences have a frame rate of 2 Hz yet the number of frames covering a certain area varies from 4 up to 18. The atmospherical conditions differ between sunny, cloudy and foggy.

For the training of our detector we have collected about 1500 object samples and the same amount of background samples as described in section 2.2. We have used 70 % of them to train a *Gentle AdaBoost* classifier with stumps as weak learners. We achived a low test error of about 7 % on the remaining 30 % of the samples. The initial number of features could be reduced from above 21000 to just 16 resulting in a fast and accurate object detector. The number of selected Haar-like features and rectangle features is nearly even. Almost all of them are extracted on the intensity channel *i1* which reveals the minor role of color information for our application.

We have manually marked the tracks of all persons in four different sequences (Fig. 4) to generate a reference for the evaluation of the detection and tracking algorithm. Altogether we have 40 frames available, each one containing about 100 up to 300 individuals. There is a large number of different evaluation metrics for object detection and tracking (Baumann et al., 2008), however we chose correctness and completeness to be sufficient for our purpose. We define a correct match between detection and reference if the distance of centroids is below 50 cm. We allow merge situations but no split situations. Tracking is evaluated by comparing all automatically generated links between consecutive images with the reference. A link is defined as correct if both associated detections match the position of the same person in the reference data.

Figure 5 displays the results of our detection algorithm as precision-recall curve. Our method is able to achieve high values for completeness and correctness yet not simultaneously. Only one test sequence stands out getting clearly better

Fig. 4. The first image of each of our four test sequences in original quality with no visual enhancement. We use the following symbols for their representation in the evaluation charts: ♦,■,▲,◄.

results. This is because it contains much less person-like clutter as the other three sequences. The results apparently show the limits of a pure detection-based approach in our domain. In many cases it is not possible to discriminate between persons and other objects solely on the base of appearance.

If we compare the detection results only with isolated persons in the reference (Fig. 5, dashed lines), the completeness increases up to 20 %. The reason for this is that our detector has been trained to find separated individuals. Groups of closely spaces persons merge to irregular shaped blobs and cannot be detected with our method.

The tracking algorithm usually fills in some missed detection and suppresses false alarms. Yet the improvement in detection is small in our case as can be seen in Fig. 6. This is because of our conservative tracking approach which allows tracks to be formed only in unambiguous situations and which prohibits multiple predictions. These restrictions could be eased to get better results. Yet at the same time a more precise motion model and a more complex data association method would be needed to account for the high object density in our domain.

The impact of the afore-mentioned problems is reflected in Fig. 7 and 8, too. Although high values are possbile for the correctness of retrieved links, the overall completeness is quiet low in all test sequences. The tracking fails especially in crowded situations when individuals are barely discriminable. Fig. 7 also shows the necessity to postpone the decision between objects and clutter from the detection to the tracking stage as it has been done in our approach. In doing so, the link completeness has been increased considerably.

The results illustrate that it is not yet feasible to use the generated tracks for a comprehensive individual behavior analysis. Yet it is possible to generate short but reliable tracklets which could be used to obtain statistics about the general motion of persons in a certain region of interest. If longer tracks are needed the presented approach needs to be extended to follow individuals even in crowded, complex situations.

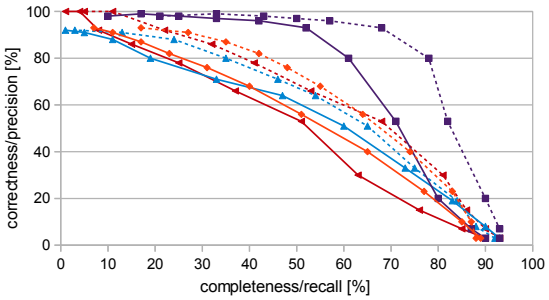

Fig. 5. Results of the detection algorithm with varying confidence threshold for four different test sequences. The evaluation is done twice, first time with the complete reference data (continuous) and second time only with isolated persons (dashed).

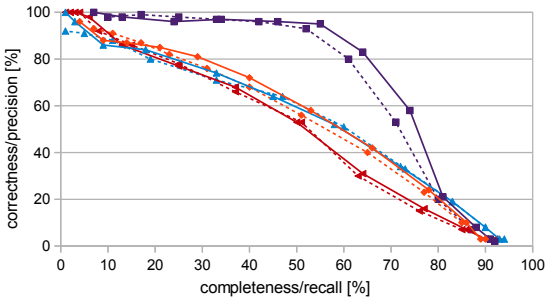

Fig. 6. Comparison of the detection results after frame-wise detection (dashed) and after tracking (continuous) for all four test sequences.

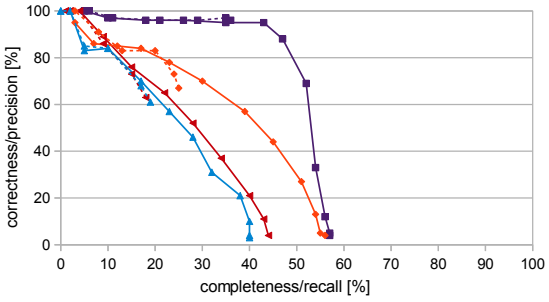

Fig. 7. Results of the tracking algorithm for four test sequences with varying confidence threshold for track confirmation. The evaluation is done twice using different detection results as input for the tracking. At first all detections with a confidence above the standard threshold have been used ($conf > 0$, dashed). Afterwards the threshold has been lowered considerably ($conf > -0.6$, continuous).

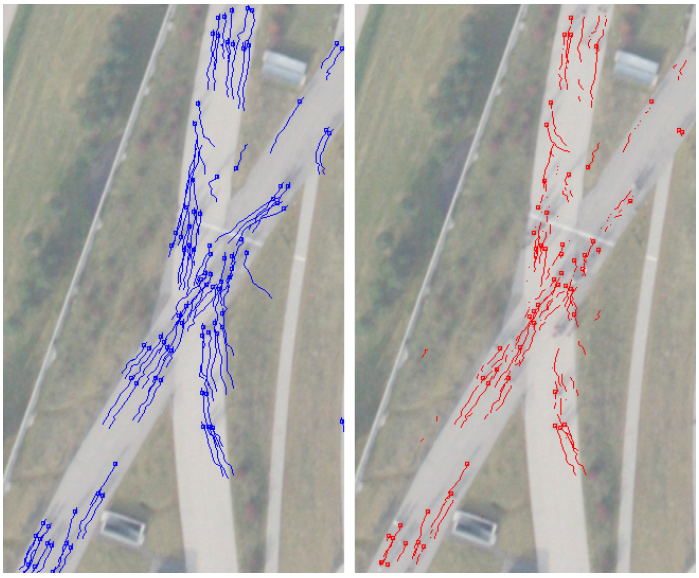

Fig. 8. Visual comparison of the reference tracks on the left and the automatically generated ones on the right for the entire sequence ■.

4 Conclusion

In this paper we proposed an appearance-based approach to detect individuals in aerial images. It comprises enhanced Haar-like features for the object shape, rectangle features for color information and a detector which is especially designed for the small size of a single person and its potential shadow. We used AdaBoost as binary classifier and demonstrated a way to collect background samples for its training very efficiently.

Our detection method achieves good results on isolated persons and in scenarios with few person-like clutter. The performance drops however in situations when people merge into groups and when a lot of clutter is present.

Furthermore we presented a tracking-by-detection algorithm which can handle a very large number of individuals and clutter in a fast way. The negative effects of a decreased detection performance are alleviated by postponing the final decision between object and background from the detection to the tracking stage. The tracking method generates short but reliable tracks of individuals which could be used to extract statistics about the general motion in a certain region of interest.

In future work these short tracks will be the base of a hierarchical tracklet-based framework in which gaps are closed and tracks are extended by the use of high level post-processing steps as in Huang et al. (2008).

References

Baumann, A., Boltz, M., Ebling, J., Koenig, M., Loos, H.S., Merkel, M., Niem, W., Warzelhan, J.K., Yu, J.: A review and comparison of measures for automatic video surveillance systems. EURASIP Journal on Image and Video Processing 2008, 30 (2008)

Benedek, C., Sziranyi, T., Kato, Z., Zerubia, J.: Detection of object motion regions in aerial image pairs with a multilayer markovian model. IEEE Transactions on Image Processing 18(10), 2303–2315 (Oct 2009)

Betke, M., Hirsh, D.E., Bagchi, A., Hristov, N.I., Makris, N.C., Kunz, T.H.: Tracking large variable numbers of objects in clutter. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2007)

Burkert, F., Schmidt, F., Butenuth, M., Hinz, S.: People tracking and trajectory interpretation in aerial image sequences. In: Photogrammetric Computer Vision and Image Analysis. IAPRS, vol. XXXVIII, Part 3A, pp. 209–214 (Sep 2010)

Collins, J., Uhlmann, J.: Efficient gating in data association with multivariate gaussian distributed states. IEEE Transactions on Aerospace and Electronic Systems 28(3), 909–916 (Jul 1992)

Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting. The Annals of Statistics 28(2), 337–374 (2000)

Grabner, H., Nguyen, T.T., Gruber, B., Bischof, H.: On-line boosting-based car detection from aerial images. ISPRS Journal of Photogrammetry and Remote Sensing 63(3), 382–396 (May 2008)

Huang, C., Wu, B., Nevatia, R.: Robust object tracking by hierarchical association of detection responses. In: Proceedings of European Conference on Computer Vision. LNCS, vol. 5303, pp. 788–801. Springer, Heidelberg (2008)

Kumar, R., Sawhney, H., Samarasekera, S., Hsu, S., Tao, H., Guo, Y., Hanna, K., Pope, A., Wildes, R., Hirvonen, D., Hansen, M., Burt, P.: Aerial video surveillance and exploitation. Proceedings of the IEEE 89(10), 1518–1539 (Oct 2001)

Leitloff, J., Hinz, S., Stilla, U.: Vehicle detection in very high resolution satellite images of city areas. IEEE Transactions on Geoscience and Remote Sensing 48(7), 2795–2806 (2010)

Medioni, G., Cohen, I., Bremond, F., Hongeng, S., Nevatia, R.: Event detection and analysis from video streams. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(8), 873–889 (2001)

Miller, A., Babenko, P., Hu, M., Shah, M.: Person tracking in uav video. In: Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007, LNCS, vol. 4625, pp. 215–220. Springer, Heidelberg (2008)

Ohta, Y.I., Kanade, T., Sakai, T.: Color information for region segmentation. Computer Graphics and Image Processing 13(3), 222–241 (1980)

Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., Poggio, T.: Pedestrian detection using wavelet templates. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 193–199 (1997)

Reilly, V., Idrees, H., Shah, M.: Detection and tracking of large number of targets in wide area surveillance. In: Proceedings of European Conference on Computer Vision. LNCS, vol. 6313, pp. 186–199. Springer, Heidelberg (2010a)

Reilly, V., Solmaz, B., Shah, M.: Geometric constraints for human detection in aerial imagery. In: Proceedings of European Conference on Computer Vision. LNCS, vol. 6316, pp. 252–265. Springer, Heidelberg (2010b)

Smal, I., Loog, M., Niessen, W., Meijering, E.: Quantitative comparison of spot detection methods in fluorescence microscopy. IEEE Transactions on Medical Imaging 29(2), 282–301 (2010)

Sung, K.K., Poggio, T.: Example-based learning for view-based human face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(1), 39–51 (1998)

Thomas, U., Rosenbaum, D., Kurz, F., Suri, S., Reinartz, P.: A new software/hardware architecture for real time image processing of wide area airborne camera images. Journal of Real-Time Image Processing 4(3), 229–244 (Aug 2008)

Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 511–518 (2001)

Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. International Journal of Computer Vision 75, 247–266 (2007)

Xiao, J., Yang, C., Han, F., Cheng, H.: Vehicle and person tracking in aerial videos. In: CLEAR: Classification of Events, Activities and Relationships. vol. 4625, pp. 203–214 (2008)

Yu, Q., Cohen, I., Medioni, G., Wu, B.: Boosted markov chain monte carlo data association for multiple target detection and tracking. In: International Conference on Pattern Recognition. vol. 2, pp. 675–678 (2006)