Image processing in aerial surveillance and reconnaissance: from pixels to understanding

Judith Dijk, Adam W.M van Eekeren, Olga Rajadell Rojas, Gertjan J. Burghouts & Klamer Schutte
TNO, Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands.

ABSTRACT

Surveillance and reconnaissance tasks are currently often performed using an airborne platform such as a UAV. The airborne platform can carry different sensors. EO/IR cameras can be used to view a certain area from above. To support the task from the sensor analyst, different image processing techniques can be applied on the data, both in real-time or for forensic applications. These algorithms aim at improving the data acquired to be able to detect objects or events and make an interpretation of those detections. There is a wide range of techniques that tackle these challenges and we group them in classes according to the goal they pursue (image enhancement, modeling the world object information, situation assessment). An overview of these different techniques and different concepts of operations for these techniques are presented in this paper.

Keywords: Cameras, Image processing, enhancement, object detection and classification, interpretation

1. INTRODUCTION

Surveillance and reconnaissance tasks are currently often performed using an airborne platform such as a UAV. Areas can be observed from above using EO/IR cameras. To support the task from the sensor analyst, different image processing techniques can be applied on the data, both in real-time or for forensic applications. In this paper an overview of such techniques and their concept of operation is presented. In the last section we also discuss the potential added value of the processing with respect to the operational task.

There is a wide range of techniques that tackle different challenges. We group the algorithms in classes according to the goal they pursue, that is image enhancement, modeling the world, object information and interpretation. The relation between these type of techniques and typical examples for the classes are presented in Figure 1-1.

The first class of techniques is image enhancement, where an improved image or series of images is produced based on the raw imagery. A challenge for airborne sensors is also motion estimation, that is to determine for the camera movement. This movement has many degrees of freedom and may have high frequency movements as well. Motion estimation is used for many image enhancement and other image processing techniques. Typical image enhancement techniques are contrast enhancement, super-resolution reconstruction and stabilization. Image enhancement techniques are the upper block in Figure 1-1.

The second class of techniques is modeling the world. Here the camera images are related to each other into one model. The camera images can also be used to make a 2D mosaic or a 3D model of the scene. These kind of techniques are the second block in Figure 1-1. For these kind of techniques a correct motion estimation is also very important.

A third type of techniques provides information about the object, such as object detection, tracking and classification. These techniques will provide answers such as "there is a moving object in the scene, this object is moving from point A to point B and it is a person". This gives the operator an idea what kind of actors are present in the scene. For airborne systems the challenge lies in finding techniques which are able to perform on small objects as well. These kind of techniques are the third block in Figure 1-1.

Electro-Optical Remote Sensing, Photonic Technologies, and Applications VII; and Military Applications in Hyperspectral Imaging and High Spatial Resolution Sensing, edited by G. W. Kamerman, O. K. Steinvall, G. J. Bishop, J. D. Gonglewski, Proc. of SPIE Vol. 8897, 88970A · © 2013 SPIE · CCC code: 0277-786X/13/$18 · doi: 10.1117/12.2029591

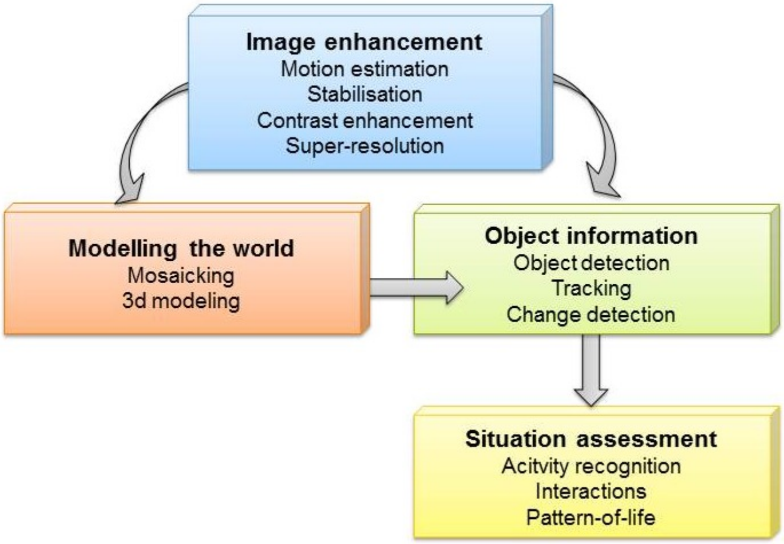

Figure 1-1 relation between different types of image processing techniques which can be applied on a UAV.

A forth type of techniques also provides interpretation by classifying actions and interactions of these actors, thereby not only answering the question who/what is presented in the scene but also what these actors are doing. A typical action is for instance running or looking at a watch. Typical interactions that can be distinguished are "two persons meeting" and "one person following another".

Interactions can be seen between two or more persons, but also for one more persons and one or more objects. This will also provide insight in the situation that is observed by the operator. These kind of techniques are presented in the lower block in Figure 1-1. For airborne systems the challenge is, as for the object detection class, that there should be enough pixels on the target to observe the action. This means that subtle actions may not be observed. All these data can be used for event detection and providing situational awareness. Here all detections and actions are used to see a certain event, and to assess what is going on. A typical long-term assessment is to determine the pattern of life. Here for instance the expected amount of persons at a specific locations and their expected movements are determined. At the start of a school day one should expect parents bringing their children to the school and then leave these premises are expected. Deviations to this patterns (so-called anomalies) may indicate a threat, e.g. an indication that a bomb is placed nearby. The challenge for airborne systems is that the time they observe a certain area is limited, but should be long enough to make these assessments.

There are other papers describing image processing for airborne platforms. Heinze et.al. [16] e.g. describe an automatic image exploitation system for small UAVs, in which many image enhancement and modeling techniques are presented. They relate the processing also to products such as geo-code image mosaics, stereo mosaics and 3D model generation, and discuss also the possibilities of data fusion with other sensors such as SAR and central databases. These techniques are mainly image enhancement and modeling the world, and contain less interpretation. Other papers like Herbin [17] discuss the image interpretation in more depth, but do not relate this to image enhancement techniques. In this paper we try to relate the type of algorithms to each other, without discussing the algorithms itself in-depth.

This paper is organized as follows. In section 2 Image Enhancement algorithms are discussed. In section 3 two modeling algorithms are presented. In section 4 we continue with object information algorithms. In section 5 the last type of algorithms, situational assessment algorithms are presented. In section 6 operational concepts and challenges for the different algorithms are discussed. The paper ends with a summary and a discussion.

2. IMAGE ENHANCEMENT

Image enhancement will improve the awareness of an operator and helps them with identification tasks. Furthermore it can improve the performance of other image processing techniques. In this section, a number of image enhancement techniques are described.

2.1 Motion estimation

Motion estimation is an enabling technique for many image enhancement techniques such as stabilization and super-resolution. The estimation of motion between images has been an active research topic for many years and a variety of methods have been developed [Zitov, 4, 6, 52]. The methods differ mainly in the type of visual features used (keypoints, lines, patches) and the motion model (e.g. Euclidean, projective or non-rigid). The visual features can be coarsely classified as belonging to either image patches or keypoints. The first type, image patches, can be used for finding similar image regions based on raw pixel values, where the search for the best matching region is often based on gradient-search[26], a variant of block matching or a measure of mutual information [47]. The local transformation that describes corresponding image patches in images is usually assumed to be a translation, although more complex transformation can also be used. For an UAV platform the motion between two frames may be too large for an accurate motion estimation on one scale. A solution for this is to use a multi-scale scheme for motion estimation [31, 42].

The second type of feature is a keypoint which is a position in the image that shows a specific recognizable structure. Examples of such keypoints are (intensity) corner points [15] or local scale-space intensity extremes [25]. At the position of the interest points, a local descriptor of the image neighborhood is computed that is largely unaffected by illumination or viewpoint changes. A frequently used descriptor is for instance the SIFT (Scale Invariant Features Transform) descriptor[24]. The comparison of the local descriptors results in a set of corresponding keypoints that describe the motion between the frames. The result after the comparison of visual features, either calculated from image patches or keypoints, is a collection of motion vectors between subsequent frames. Depending on the motion of the camera w.r.t. some static scene we can identify a global transformation of all the vectors. In particular, when the camera movement does not include any translation but only rotation about its center, a good approximation for the global transformation is a projective transformation. Another possibility for the approximation by a projective transformation to be valid, is that all scene points lie on a plane. In practice this will hold approximately for distant scenes. In a static world with a non-static UAV the projective transformation, or its approximation by an affine transformation, will be adequate for describing the frame-to-frame motion.

The merit of calculating a global transformation for the motion is that it can be used to correct noisy and also erroneous motion vectors, the so-called outliers. These are caused by areas showing little image structure so that the underlying movement cannot be determined. Also when the assumed motion model for the segment is not correct, such outlying motion vectors may appear. Robust estimation methods, like [9, 19] are therefore applied for the estimation of the global transformation.

2.2 Stabilization

On small platforms, such as a tactical UAV, stabilization of an optical sensor is very important for visual quality. Stabilization can be obtained with a mechanical construction, but this will not compensate for all perturbations. Advanced image enhancement techniques are a cost- and weight effective alternative for mechanical stabilization. But they can also help to improve the stabilization result of mechanical stabilization.

Stabilization can be done by correcting for the motion that was measured with the motion estimation algorithms described in the previous section. The challenge is to correct only for the vibrations and unwanted motions, and not for the global translation of the UAV.

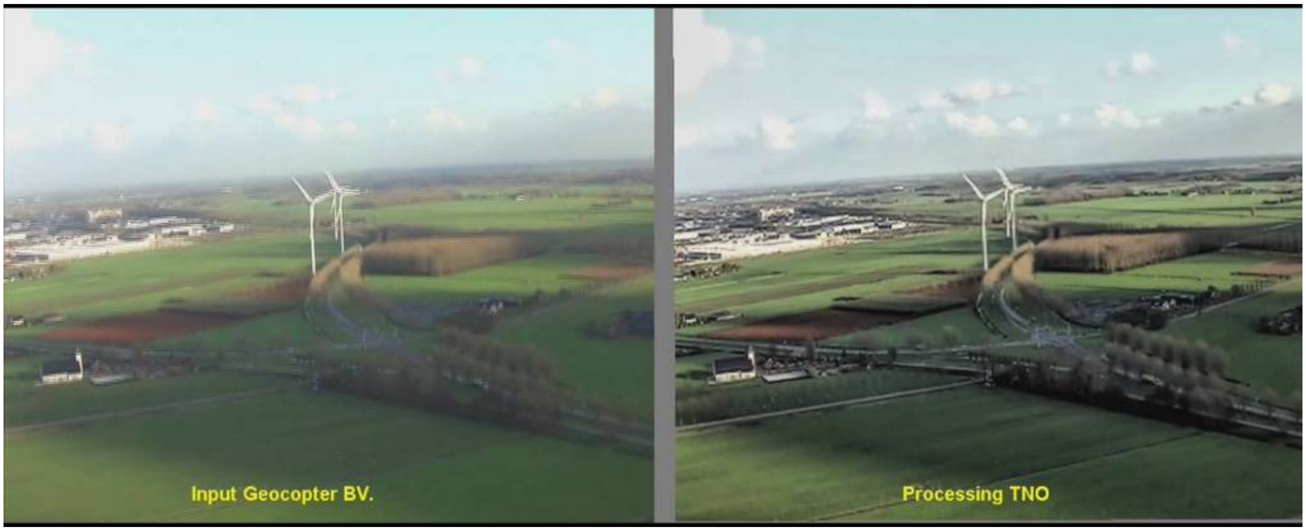

Figure 2-1: Left: one of the input images taken from an unmanned helicopter (Geocopter BV). Right: result after processing with motion compensation, super-resolution, stabilization and contrast enhancement.

2.3 Contrast Enhancement

To enhance the contrast in an image several methods are described in literature [27, 28, 30, 35, 41, 48, 51] The most simple way to enhance the contrast is by adjusting the image to the available range, so-called global contrast stretching. This will help in cases were only part of the available range is used, but will fail in situations where the range used in one part of the image is very different than in another. To enhance the contrast in a part of the available contrast range gamma manipulation can be done. In this case details in part of the luminance range of the image will be enhanced, at the cost of decreasing the luminance in other parts of the image. Another global method is histogram equalization, were the luminance values are changed so that the histogram of the image are as flat as possible. The main disadvantage of this method is that the output images are not so natural anymore.

To enhance local contrasts grey-value local adaptive contrast enhancement can be used. The idea of local contrast enhancement is that the processing depends features in a local region [30, 41]. An example of local contrast enhancement is presented in Figure 2-1. It can be seen that e.g. the trees are better visible after processing.

2.4 Super resolution reconstruction

Super-Resolution is an image enhancement technique that improves the spatial resolution of images by using temporal information[36, 42]. The input is a video sequence (e.g. 30 frames). If the displacement between all frames is estimated at sub-pixel accuracy, all frames can be registered and mapped to a reference frame with a denser pixel grid than that of the original frame. In such a way we have obtained an image of the same scene with a higher resolution. This can also be done for every frame, in which case the result is a video instead of a static image.

For tactical purposes people are especially interested to enhance the visual quality of specific objects of interest, such as vehicles, buildings and persons. Most research is performed on enhancement of the background and rigid objects (moving vehicles) because these kinds of objects are easier to process. Moving non-rigid objects, such as persons, are very difficult to register with sub-pixel accuracy. An example of super-resolution is shown in Figure 2-1 and in more detail in Figure 2-2 . It can e.g. be seen that the wind mill is sharper in the processed image.

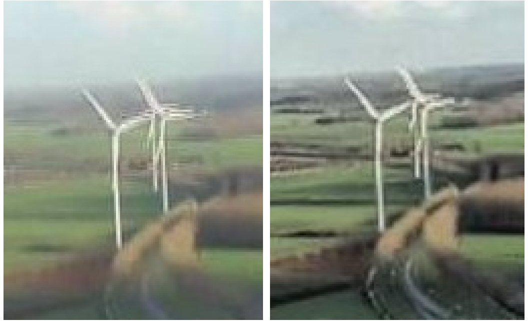

Figure 2-2 Zoom in into the results of Figure 2-1.

3. MODELING THE WORLD

The second class of techniques is modeling the world. Here the camera images are related to each other into one model.

3.1 Mosaicking

The idea of mosaicking is that the different images are related to each other and combined in one single overview image. The first step is performing motion estimation. After that the images need to be overlaid or fused. Moving objects in the imagery can appear on different locations in the footage, and therefore need to be treated differently from the static background. Mosaicking works best for images for which a 2D world model can be assumed. An example of image mosaicking is shown below. The different images can still be seen by looking at the corners in the images.

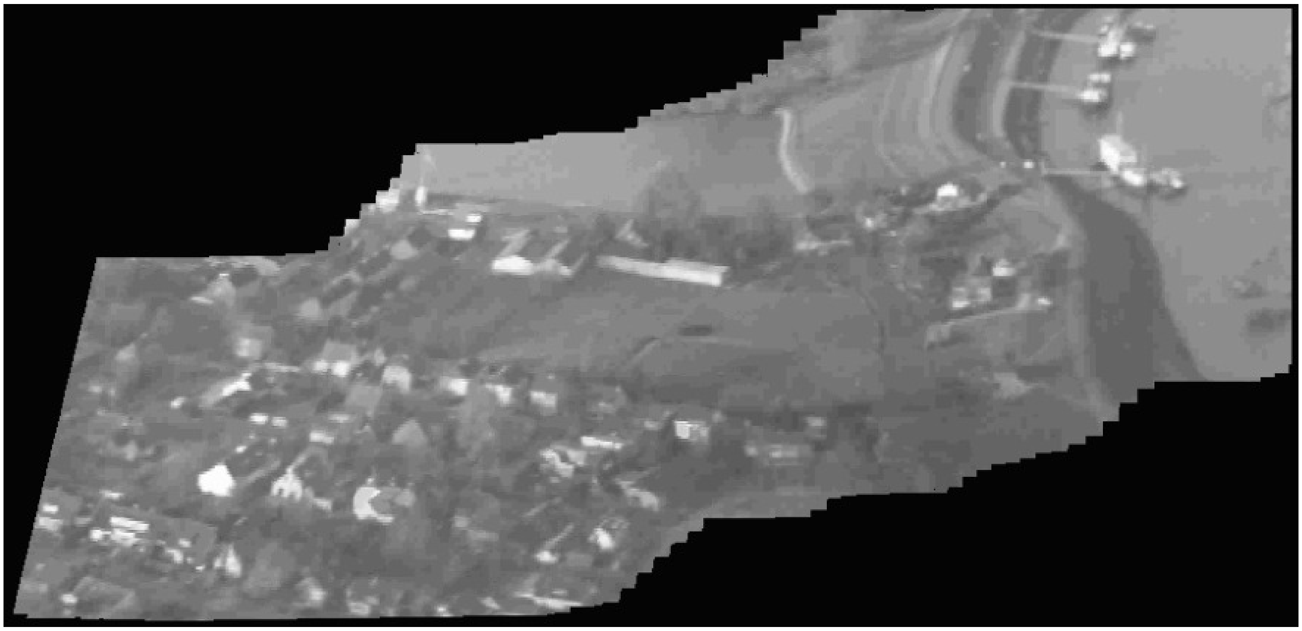

Figure 3-1 Example of mosaicking

3.2 3d modeling

3D modeling is a technique to extract 3D information from multiple 2D images. When the earth surface is observed from the sky this 3D information reveals the height of objects on the ground. This can be used for visualization of the environment, and in further steps for tasks as object detection and classification, obstacle detection and change detection. 3D modeling also depends heavily on a correct motion estimation. An example of typical input data is shown in Figure 3-2.

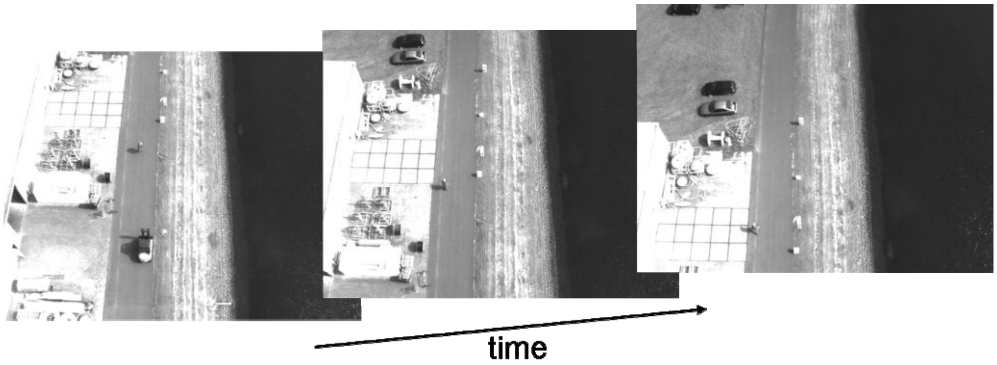

Figure 3-2 Three out of 33 captured images from an unmanned helicopter flying at approximately 50m altitude. A piles of boxes are visible at the right side of the road

After motion estimation all camera positions and orientations are determined and can be plotted in one coordinate system. For the example, this visualization is presented in Figure 3-3. After the egomotion estimation dense 3D point cloud is estimated using the software package PMVS [10]. This result is also presented in Figure 3-3.

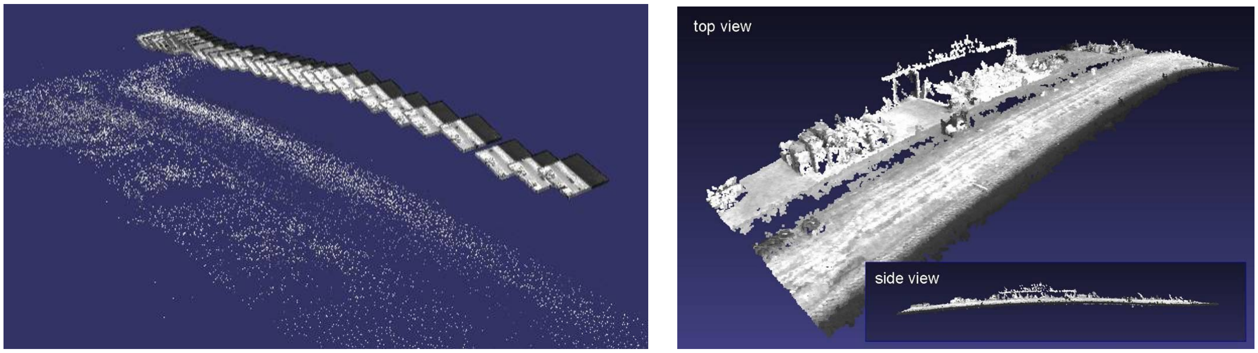

Figure 3-3 Visualization of all estimated camera positions and orientations (left). The white dots are estimated 3D points of the earth surface. Top and side view of 3D point cloud of the reconstructed scene (right)

4. OBJECT INFORMATION

The third class of processing algorithms provided information about the objects in the image. Algorithms in this class will return information about the number and type of objects in the imagery. Object detection algorithms are discussed in section 4.1. If the objects are detected, they can be tracked (section 4.2). A special type of object detection is change detection (section 4.3), where changes in a certain scene recorded at different times are detected.

4.1 Object detection

Object detection can be done in several different ways. Objects can be found using spatial features in single frames, or by using temporal features such as motion. Next to that, one can look for dedicated objects, or for general objects. The search for dedicated objects also results in classification, that is the object are not only detected but also recognized as being part of a certain class. It is also possible to use knowledge about the scene in the detection, for instance the knowledge that a car is normally seen at a road. This can be done with user input [12] or insupervised [43]. Of course also different types can be combined [44]. In this paper we discuss two classes of object detection in more detail: generic moving object detection and dedicated static object detection. An example of generic static object detection is Objectness [1], in which a scene is classified in different objects which are not known on forehand. An example of Dedicated moving object detection is detection with motion templates [29], where a specific motion is used as search pattern. These algorithms are not discussed in detail in this paper.

4.1.1 Generic moving object detection

The idea of generic moving object detection is that if the background can be substracted from an image, only the moving objects remain. For this the imagery needs to be registered using the algorithms discussed in section 2.1.

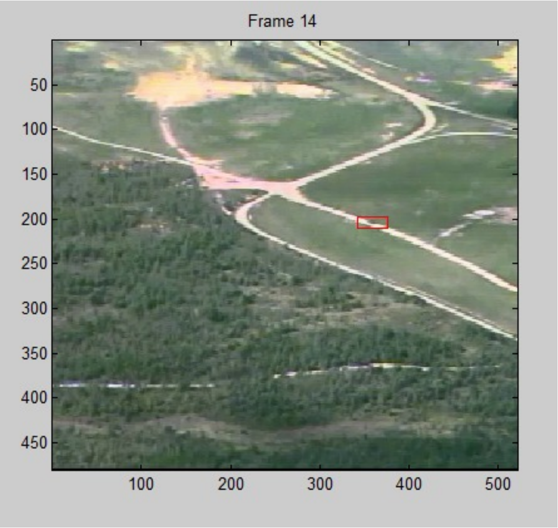

Figure 4-1 Example moving object detection based on motion

The performance of state-of-the-art moving object detection methods is very good and can be used in real-time applications.

The main advantage of these type of techniques is that they can be applied for small objects, even for single pixel objects. Such an object and detection is shown in Figure 4-1. The main drawbacks of this method is that is heavily depends on the quality of the motion estimation, and that 3D objects will generate false detections.

4.1.2 Static specific object detection

In static specific object detection objects such as persons and vehicles are detected from single image frames. In the literature there are various methods available for specific object detection. We discuss three of these detection methods in more detail: SIFT, Felzenschwalb and Viola-Jones.

Lowe introduced a keypoint based detection system using Scale Invariant Feature Transform (SIFT) descriptors [24]. In an image a large number of keypoints (typically hundreds of keypoints) is detected and the keypoints are then compared to the keypoints of known objects. Many variations of this method exist with different detectors, different matching techniques and different clustering or scoring functions. The method is very robust to small viewpoint changes and is often robust to variations in the pose of an object. The main drawback is that the objects to be detected must be large enough to have a sufficient number of robust keypoint. Objects of only a few pixels will not be detected.

The Felzenszwalb detector [7] is a highly successful detector for generic deformable object types, such as humans. For a large set of example images a model is trained that consists of deformable parts. The different parts of the model make this detector robust against small changes in pose of the object. The added complexity of the model is reflected in the fact that the detector needs a large amount of training data and high computation time.

The Viola-Jones detector [46] was introduced in 2001 as a real-time face detector. As this method is generic it can be trained for any class of objects. Disadvantages of the method are the fact that the method does not handle non-rigid models (such as humans) very well because the differences between individuals are countless. Consequently, different viewpoints of the same object are not supported. However, this can be an advantage when dealing with small objects were less detail is available and non-rigid models become almost rigid.

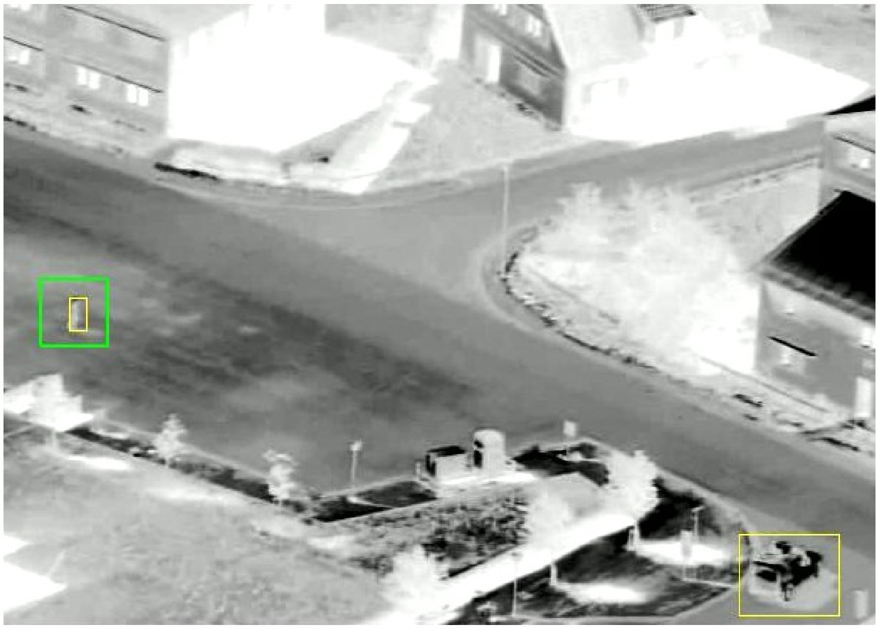

Figure 4-2 Example object detection, based on motion (yellow) and person detection (green)

In Figure 4-2 object detections in IR imagery are shown, with detection based on motion (in yellow) and on the Viola Jones detector trained on humans (in green). In Figure 4-3 three different zoom levels for aerial video data are shown. The size of the car in the bottom of the images on the left, center and right is about 16, 50 and 140 pixels, respectively. This object size is related to the detection probability. Especially the static object detectors require a minimum object size for detection. In the data shown in Figure 4-2 we were able to detect persons of 14 pixels length. On the UCF data shown later for activity recognition the minimum height for which a person could be reliably detected was about 40 pixels. Below that the thresholds on a person where such that there were either a large number of misses, or a low detection rate.

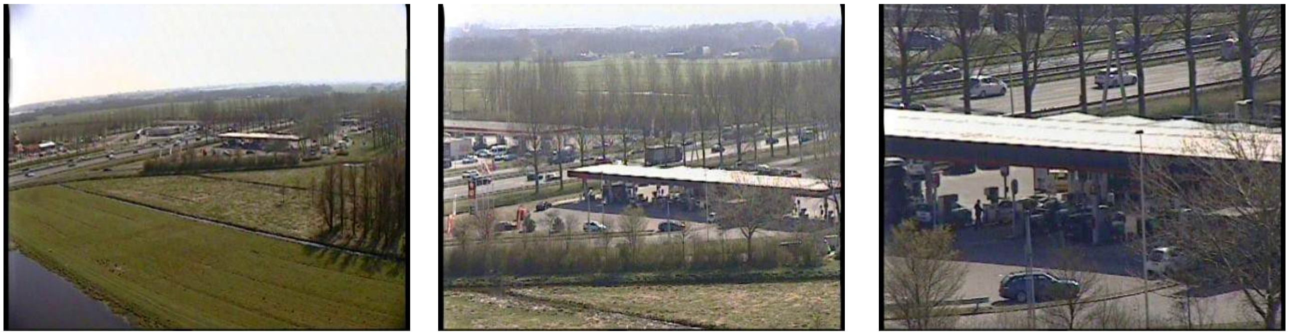

Figure 4-3 Three different zoom levels for aerial dataset.

4.2 Object tracking

Objects can be tracked throughout the video sequence by making use of their position, movement and possibly appearance. Tracking will provide the association of multiple detections that belong to a single entity, and will describe the path that the entity has followed in the scene. The amount of literature on tracking methods is vast; a recent overview of methods for person tracking is given in the Performance Evaluation of Tracking and Surveillance Workshop [8]. Note that this tracking can also be done on generic or dedicated features.

Figure 4-4 and Figure 4-5 show results of a generic object tracker applied to a sequence of the UCF Aerial dataset [18]. This sequence has been recorded with a 960x540 pixel HD TV camera at an altitude of approximately 100m. In Figure 4-4 the moving objects in the scene are all tracked, but track consistency is lost when one of the persons goes out of view due to the jerky camera movement. In order to remedy this problem, a track history should be kept that tries to match new tracks to historical tracks in order to keep the same track number. In Figure 4-5 the tracks over a longer time span are shown, indicating that track numbers can be preserved once sufficient (> 10) detections (due to movement) are found. Note that parallax effects of static objects can result in false alarms (see track nr. 313 in bottom frame).

There exist many different methods for object detection and tracking. Most of these methods perform very well on high-resolution data where the objects cover hundreds of pixels. In aerial data objects are often very small, covering less than hundred pixels. Detection of such small static objects (when they are not moving) is a very challenging task in color video, even in good weather conditions. For IR data the detection might be better, which will also result in a better tracking.

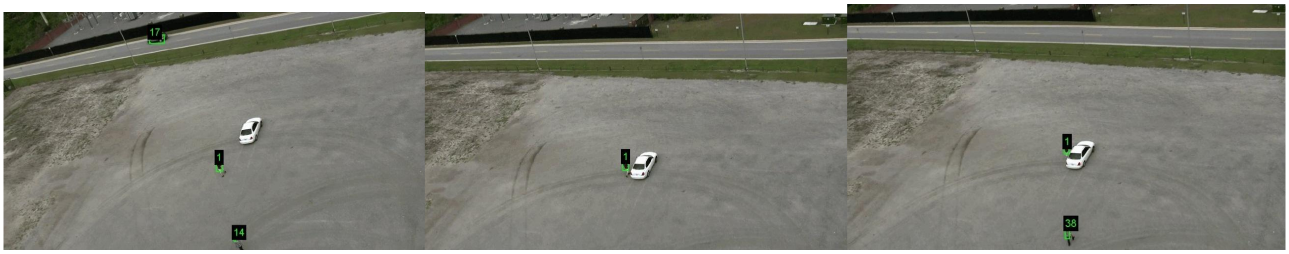

Figure 4-4 Frame numbers 30, 50 and 56 (top to bottom) in an UAV sequence. Due to the camera shake one of the persons goes out of view (track nr. 14), and reinitiates a track when he’s back in view (track nr. 38).

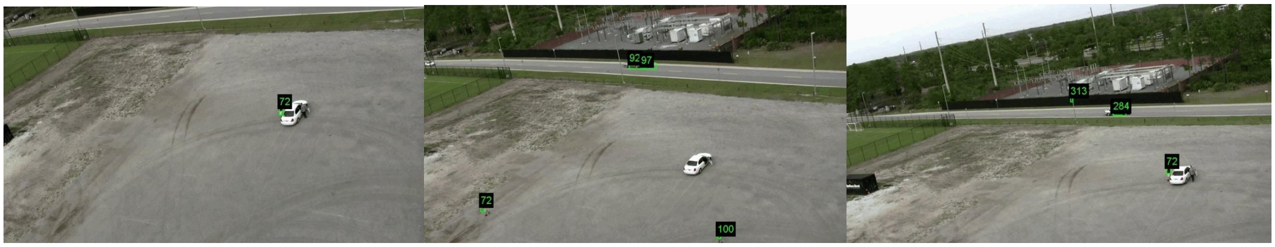

Figure 4-5 Tracking results for frames 125, 160 and 270 (top to bottom) in the same sequence as in Figure 3.11. Here it is shown that a person (track nr. 72) can be tracked over a long period. Due to the fact that the person remained stationary after frame 56 (see Figure 3.11) the track for this person (track nr. 1) was lost due to missing detections.

It is also possible to track a certain object which is detected in a single frame or pointed out by an operator. In Figure 4-6 an example is shown where an unmanned ground vehicle is followed within the UAV camera footage. In the first frame a rectangle is drawn around the UGV by the operator of the UAV camera, after which the processing takes over. Here the detection is based on a dedicated template, but the tracking is generic. The tracking information can also be fed into the UAV system (both camera and vehicle) so that the UGV will always be in the center of the image.

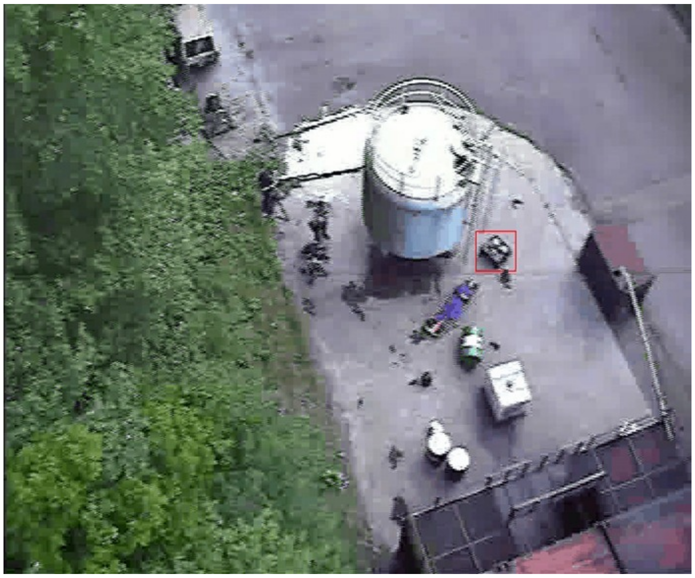

Figure 4-6 Object tracking based on a template

4.3 Change detection

There exist many ways to perform 2D change detection in image sequences. An important pre-processing step is to align the images in both sequences. For this alignment it is necessary that we can assume that the scene is 2D, otherwise the parallax effect will introduce distortions that are falsely identified as changes. We are interested in local change detection, since small changes such as the appearance or disappearance of vehicles and individual persons can be of interest.

When the scene is approximately flat, the best way to perform change detection is by comparing pixel differences. An intensity correction is necessary to compensate for lightning changes. Here an example based on footage of the daylight data recorded from a Sperwer UAV is shown. A typical pair of images from a run with and without object is shown in Figure 4-7. The result after geometric correction is shown in Figure 4-8. It can be seen that there are still a lot of changes found on differences in vegetation. When the scene is not flat (like in this case) 2D change detection will probably result in too much false alarms. For change detection of such scenes it is better to perform the change detection in 3D.

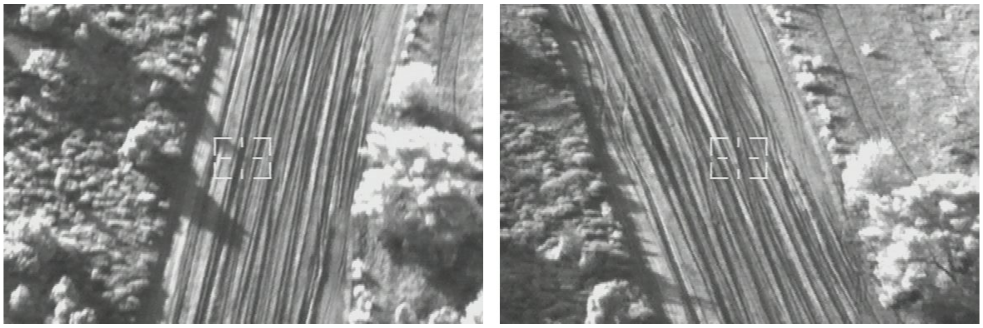

Figure 4-7 Daylight data from the Sperwer from two different runs.

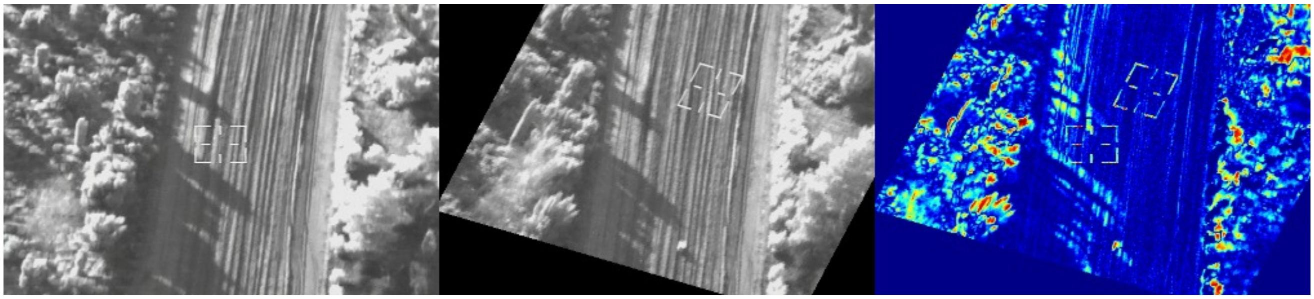

Figure 4-8 Daylight data after geometrical correction and differences found.

An example of 3D change detection is shown in Figure 4-9. At the left an image with cars, and at the right the same position without cars is given. The different in height is show below, both before and after threshold. It can be seen that the two cars, i.e. the changes, stick out in the difference image, and can therefore easily be detected using a threshold. Here the 3D information is used for the change detection.

An overview of change detection techniques is presented by Radke [37].

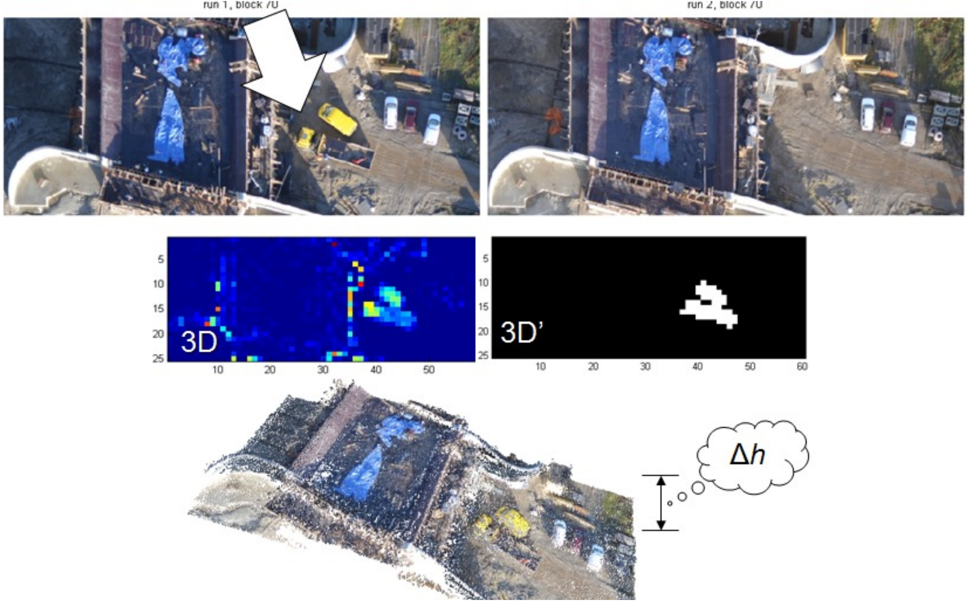

Figure 4-9 Example of 3D change detection

5. SITUATION ASSESSMENT

5.1 Activity recognition

Recognizing human actions is a critical aspect of many types of surveillance. Human actions can be indicative of a wide range of unwanted situations such as aggression, vandalism, theft, somebody falling, and becoming unwell. Recognizing actions is an active field of research [2, 11, 14, 18, 21] and promising results have been shown on explicit actions such as aggression detection [23], simple kinematic actions such as walk, bend and jump [13, 39] and sports [38]. More complex actions that involve subtle motions, for instance give and put down, are still not yet well recognized [3]. A nice overview of scene understanding algorithms for Aerospace sensors is also provided by Herbin et.al [17]

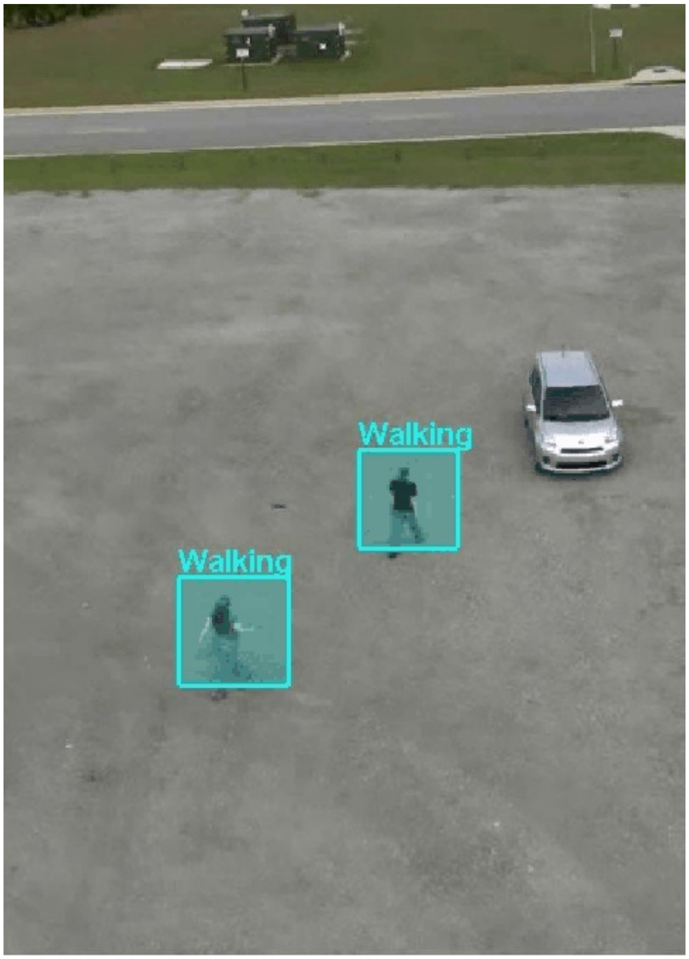

Figure 5-1 activity recognition

To detect different actions first the humans within the scene need to be detected and tracked. In our work, we use small tracks (tracklets) of approximately five to twenty frames. These small tracklets are large enough to provide activity classifications, even close to vehicles.

For the actual action recognition the bag-of-words (BoW) model, based on low-level motion features, is a frequently used model. This is a simple model that has achieved good performance on a range of action recognition tasks, e.g [49]. In many applications STIP features [22] are used as input features because they proved to be robust features for describing actions [3, 21, 23]. These features are hard to use for UAV data because the algorithm uses the assumption that the camera is either completely static or perfectly stabilized, which is hard for a camera mounted on a free moving platform. One of the challenges that we identified is to make a robust version of the STIP which can work with a small registration error.

An example of activity recognition is shown in Figure 5-1. Instead of STIPS we used a tracklet with a length of 5 frames in a fixed cube of 20x20x5 (height x width x time). In this cube we extracted the following features

- Local binary patterns [32],
- Standard deviation in the time direction
- Gradient in the spatial directions
- Raw pixel values
- Location of all features in the cube.

Classification is done with the Bag-of-Words representation. Here it can be seen that the activity of the two persons in the scene is classified as walking, which is indeed correct.

The main difference for UAV activity recognition compared with recognition in many other camera systems is that the persons are observed from above, which provides different imagery than from the side. The size of the persons can be very small, which limits the detection of more subtle actions. Also the time that a specific area and/or specific persons can be observed within the camera footage is limited.

5.2 Interactions

Instead of only looking at actions of one person, one can also interpret the interaction in the scene, Interactions can be seen between two or more persons, but also for one more persons and one or more objects. These type of techniques not only answer the question what a certain actor is doing, but also how this relates to each other. A typical action is for instance running, where a typical interaction is one person chasing another or a person getting out of a vehicle. An example of the latter interaction is shown in Figure 5-2. Note that the quality of both the activity recognition and the interaction detection depends heavily on the quality of the person detection and tracking.

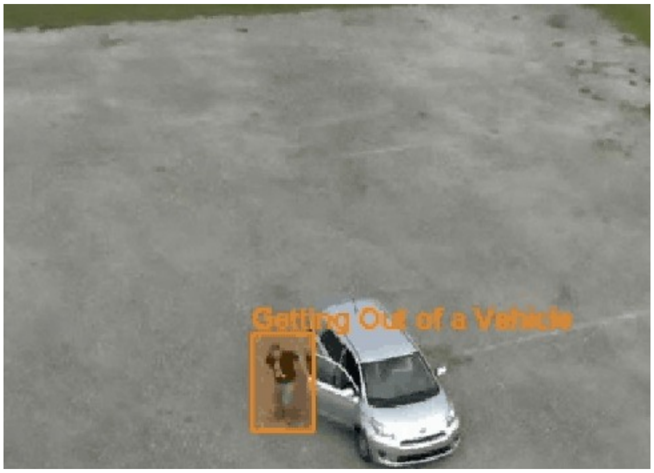

Figure 5-2 Example of automatic detection of a person getting out of a vehicle.

5.3 Pattern-of-life analysis

Action and interaction detection data can be used for event detection and providing situational awareness. For instance, the detected objects and actions are used to see a certain event, and to assess what is going on. Also the detection of pattern-of-life patterns and anomalies to this patterns can be determined.

Most people and populations tend to have very distinct patterns in their lives. Pattern-of-life analysis [34] works with the assumption, that if behavior patterns of e.g. a population, town or street are continuously observed, the regular patters can be identified and deviations (or anomalies) to this pattern can be detected. For instance, if it is nice weather on a holiday and no children are playing in the streets, that is unexpected. The algorithms shown before can support pattern-of-life analysis and anomaly detection on this type of data. The challenge for airborne systems is that the time they observe a certain area is limited, which makes it challenging to determine the normal pattern.

6. OPERATIONAL CONCEPTS & CHALLENGES

In the previous sections we provided an overview of different image processing techniques for UAV data. All these algorithms can be used to support the operator, either by providing a better image or by providing more information about the data such as object detections, classifications or actions. In this section some thoughts on operational concepts and challenges with respect to the use of these techniques are given.

Image enhancement algorithms provide higher quality images. These images can be used as input for further processing and/or can be observed by humans to interpret the scene. The drawback of the latter is that the observation of a video of a certain time usually costs at least a similar time as the observation itself. This is mainly a problem for surveillance and searching. For tasks like reconnaissance and battle damage assessment, this is a smaller problem as the location and time frame of the observation can be much smaller. Image enhancement techniques will help here to see details in the imagery better and faster. This holds both for real-time, more tactical as for offline applications such as forensic and intelligence tasks. The challenge for image enhancement is to improve the visibility of details without introducing (too much) artifacts. Many image enhancement techniques are relying on a well performing estimation of the camera motion. A challenge here is to determine what motion is actually caused by moving the camera and what motion by moving objects within the scene.

Techniques like mosaicking and 3D modeling can be used to make an overview of some video footage and relate the different images to each other. This will reduce the time needed for the operator, and will enable the relation of different objects that are located further away from each other. The challenge for these algorithms is to relate all data from the same objects to each other, so that no duplicates of these structures will appear. 3d modeling is also a good pre-processing step for other algorithms such as object detection, classification and change detection.

The results of object information algorithms will both support surveillance and searching, as reconnaissance tasks. For real-time surveillance tasks, the object detection will be the main benefit. This will make the detection of objects of interest within the data much faster. This can be supported by the other algorithms. For instance, if one is interested in persons and movements of persons, it will really help if a trigger is given for all persons that can be seen in a scene. If such a trigger distinguishes between humans and dogs that can be seen, that will hugely help. The camera can be zoomed on detections so that the persons are seen in higher detail. This can also be used for gathering intelligence data about the persons more quickly.

When the data is already collected and viewed afterwards, such a zoom or cue cannot be done. For some applications it would greatly help to collect all data containing specific information, such as all data with a human or a group of humans in it. This search can also be more specific, such as all blue cars (or a specific blue car) within a certain area and time frame. A huge amount of time is saved if such a search can be performed automatically or semi-automatically. Trinh [45] presented an approach in which the UAV image data was summarized into small clips with objects detections and tracks. He showed on the Virat dataset an impressive data reduction rate.

One challenge here is to have the information reliable enough for the task at hand. A second challenge is to relate the information need of a person to specific processing, and to perform this processing for the most part before the question was asked. A project aiming to provide the capability to search semantically for any relevant information within "all" (including imaging) sensor streams is the GOOSE project [40]. A challenge related to object information algorithms on UAV data is to find small persons with little detail within the imagery. There exists many techniques for object detection but few that are effective when the objects are only a few pixels size.

The activity recognition and situational awareness algorithms are still in a research phase. These algorithms can be useful for real-time tasks and for off-line, more intelligence tasks. In real-time tasks such as surveillance and reconnaissance tasks, the added value is seen especially when the decision time is crucial, and/or for situations in which anomalies with respect to the expected pattern-of-life can be seen. For intelligence tasks and for forensic application these algorithms enable a more extensive search with more different parameters. Also the detail in which automatically a pattern-of-life within a specific area can be determined is more detailed. The first challenge for these type of techniques is to find out what the possibilities for UAV data are. After that, the requirements for specific use cases can be determined.

7. SUMMARY AND DISCUSSION

In summary, the task of the UAV sensor analyst can be supported by different image processing techniques both in real-time or for forensic applications, Some image processing techniques such as image enhancement can be used as pre-processing step for other techniques. The added value of different algorithms depends also on the task of the operator and the situation in the scene that is observed. There are three different subjects related to image processing that we did not discuss in detail in this paper. These are related to new camera systems, feedback between vision and UAV and evaluation by real operators

In this paper we only evaluated normal cameras. A new development are Wide-field-of-View cameras, which can be used to observe an area of a number of squared kilometers continuously. Processing these cameras has its own challenges and benefits. An example of processing for these type of cameras is presented in [50].

The results of the image processing algorithms can also be used to for flight automation of an UAV. The tracking of the object, shown in Figure 4-6 can for instance be used to keep the object automatically in the center of the video, by automatically steering the camera to this point and/or changing the flight path of the UAV, in other wordt, real-time flight control. These kind of techniques are also presented in [31].

Many techniques presented in this paper are not evaluated extensively by military operators yet. These experiments are planned in the near future.

ACKNOWLEDGEMENTS

The work for this paper was partly done within the Unmanned Systems programme (V1340) in which TNO researches the use of different unmanned systems.

REFERENCES

[1] Alexe, B. . Thomas Deselaers, and Vittorio Ferrari. "Measuring the objectness of image windows," (2012)

[2] Burghouts, G.J., Eendebak, P., Bouma, H. , ten Hove, J..M., “Improved action recognition by combining multiple 2D views in the bag-of-words model,” AVSS, (2013).

[3] Burghouts, G. J. Schutte, K., “Spatio-temporal layout of human actions for improved bag-of-words action detection, “ PRL, (2013).

[4] Dawn, S., Saxena, V. and Sharma, B., , “Remote Sensing Image Registration Techniques: A Survey," Lecture Notes in Computer Science 6134, ICISP 2010, pp 103-112 (2010)

[5] van Eekeren, A.W.M, “Capabilities of electro-optical sensors for short-range tactical UAVs,” TNO-DV report, A429, (2010).

[6] Eendebak, P.T., van Eekeren, A.W.M., den Hollander, R.J.M. “Landing spot selection for UAV emergency landing,” *SPIE Defense Security and Sensing*. International Society for Optics and Photonics, (2013)

[7] Felzenszwalb, P.F., R. B. Girshick, D. McAllester, and D. Ramanan, “Object Detection with Discriminatively Trained Part-Based Models,” PAMI, vol. 32, no. 9, pp. 1627 –1645, (2010).

[8] Ferryman, J.L. [Proc. of 11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance], (2009).

[9] Fischler, M. A. and Bolles, R. C., “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Commun. ACM, 24, (1981).

[10] Furukawa Y. and J. Ponce, “Accurate, Dense, and Robust Multi-View Stereopsis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, Issue 8, (2010).

[11] Gorelick,L., M. Blank, E. Shechtmanm, M. Irani, R. Basri. “Actions as space-time shapes,” PAMI, (2007).

[12] Guilmart, Christophe, Stéphane Herbin, and Patrick Pérez. "Context-driven moving object detection in aerial scenes with user input," ICIP,, (2011).

[13] Guha,T., R.K. Ward, “Learning sparse representations for human action recognition,” PAMI, (2012).

[14] Gupta A., , P. Srinivasan, J. Shi, L. S. Davis. “Understanding videos, constructing plots: learning a visually grounded storyline model from annotated videos,” CVPR, (2009).

[15] Harris, C. and Stephens, M., “A combined corner and edge detector," In Proc. of Fourth Alvey Vision Conference, (1988).

[16] Heinze, N., et al. "Automatic image exploitation system for small UAVs." SPIE Defense and Security Symposium. International Society for Optics and Photonics, (2008).

[17] Herbin, S., F. Champagnat, J. Israel, F. Janez, B. Le Saux, V. Leung, A. Michel, "Scene Understanding from Aerospace Sensors: What can be Expected?," AerospaceLace, issue 4, (2012)

[18] Ikizler-Cinbis, N., S. Sclaroff. “Object, scene and actions: combining multiple features for human action recognition,” ECCV, (2010).

[19] Jackson, Brian P., and A. Ardeshir Goshtasby. "Registering aerial video images using the projective constraint," IEEE Transactions on Image Processing, (2010)

[21] Laptev,I., M. Marszalek, C. Schmid, B. Rozenfeld. “Learning realistic human actions from movies,” CVPR, (2008)

[22] Laptev, I., “On space-time interest points,” IJCV, 64 (2/3), (2005).

[23] Lefter, I., L.J.M. Rothkrantz, G.J. Burghouts, “A comparative study on automatic audio-visual fusion for aggression detection using meta-information, “ PRL, (2013).

[24] Lowe, D.G. “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal of Computer Vision, vol. 60, pp. 91–110, (2004).

[25] Lowe, D. G., “Object recognition from local scale-invariant features," in Proc. of the International Conference on Computer Vision, (1999).

[26] Lucas, B. D. and Kanade, T., “An iterative image registration technique with an application to stereo vision," *IJCAI*. Vol. 81, (1981).

[27] Moroney, M., “Local Color Correction using Non-Linear Masking.” IS&T/SID Eight Color Imaging Conference, pp 108-111, (2000)

[28] Mukherjee, D. and B. Chatterji, “Adaptive Neighborhood extended contrast enhancement and its modifications,” Pattern Recognition Letters, vol. 11, pp. 735-742, (1990)

[29] Müller, M., and Röder, T., "Motion templates for automatic classification and retrieval of motion capture data," Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation, Eurographics Association, (2006).

[30] Narenda, P.M. and R.C. Finch, “Real-time adaptive contrast enhancement,” IEEE transactions on pattern analysis and machine intelligence, vol. 3, no. 6, pp. 655-661, (1981)

[31] Nordberg, Klas, et al. "Vision for a UAV helicopter," IROS, Lausanne, Switzerland (2002).

[32] Ojala,T, M. Pietikäinen, and T. T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary pattern,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, (2002).

[33] Oh, S., A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, “A large-scale benchmark dataset for event recognition in surveillance video,” CVPR, , pp. 3153 –3160. (2011)

[34] Mak.com, “http://www.mak.com/bloghome/tag/-pattern-of-life-analysis.html,” (2013)

[35] Paranjape, R.B. , W.M. Morrow and R.M. Rangayyan “Adaptive - neighborhood histogram equalization for image enhancement Computer Vision,” Graphics and Image Processing: Graphical models and image processing, vol. 52, no. 3, pp. 259-267, (1992)

[36] Park, S.C., M.K. Park, and M.G. Kang, “Super-resolution image reconstruction: a technical overview,” IEEE Signal Processing Magazine, vol 20, no 3,, (2003)

[37] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, “Image Change Detection Algorithms: A Systematic Survey,” IEEE Transactions on Image Processing, vol. 14, pp. 294–307, (2005).

[38] S. Sadanand, J. J. Corso. “Action bank: a high-level representation of activity in video,” CVPR, (2012).

[39] C. Schuldt, I. Laptev, B. Caputo. “Recognizing human actions: a local SVM approach,” ICPR, (2004).

[40] Schutte K., Bomhof, F., Burghouts, G.J.J. , van Diggelen, J. ; Hiemstra, P. et al., " GOOSE: semantic search on internet connected sensors, ", Proc. SPIE 8758, Next-Generation Analyst, (2013)

[41] Schutte, K., “Multi-scale adaptive Gain control of IR images.” SPIE,vol. 3061, pp. 906-914, (1997)

[42] Schutte,K., D.J.J. de Lange, S.P. van den Broek, “Signal conditioning algorithms for enhanced tactical sensor imagery,” SPIE,, vol 5076, pp92-100, (2003)

[43] Shi, Xinchu, et al. "Context-driven moving vehicle detection in wide area motion imagery," Pattern Recognition (ICPR), (2012)

[44] Teutsch, M., Krüger, W. and Heinze, N.,"Detection and classification of moving objects from UAVs with optical sensors," SPIE Defense, Security, and Sensing. International Society for Optics and Photonics, (2011).

[45] Trinh, H. et al., "Efficient UAV video event summarization," ICPR, (2012).

[46] Viola P. and Jones,M., “Robust Real-time Object Detection,” (2001).

[47] Viola, P., and William M. Wells III. "Alignment by maximization of mutual information," International journal of computer vision, 24.2, (1997).

[48] De Vries, F. “Automatic, adaptive, brightness independent contrast enhancement,” Signal Processing, vol. 12, pp. 169-182, (1990)

[49] X. Wu, D. Xu, L. Duan, J. Luo. “Action recognition using context and appearance distribution features,“ CVPR, (2011).

[50] Xiao, Jiangjian, et al, "Vehicle detection and tracking in wide field-of-view aerial video," CVPR, (2010).

[51] Young, I.T., J.J. Gerbrands, and L.J. van Vliet, “Image Processing Fundamentals,” Madisetti, V.K. and B.D. Williams, editors, The Digital Signal Processing Handbook, chapter 51, pages 1-81. IEEE Press and CRC Press, (1998)

[52] Zitov, B. and Flusser, J., “Image registration methods: a survey," Image and Vision Computing, 21, (2003)

APPENDIX A: TEST DATA

There exist several datasets which can be used for evaluating image processing algorithms for aerial data. In this section the public datasets used in this paper are given.

UCF: http://server.cs.ucf.edu/~vision/aerial/index.html.

This UCF dataset contains annotated moving persons and cars, of considerable size (approximately 40-50 pixels length). The platform altitude is about 100-150 meter. The motion is random.

VIRAT [33]

The Virat dataset contains smaller moving persons and cars, of about 8-15 pixels length. The platform altitude is about 500-1000 meter. The motion is more uncontrolled than UCF, and also contains panning. A drawback of the dataset is that the data is not annotated.

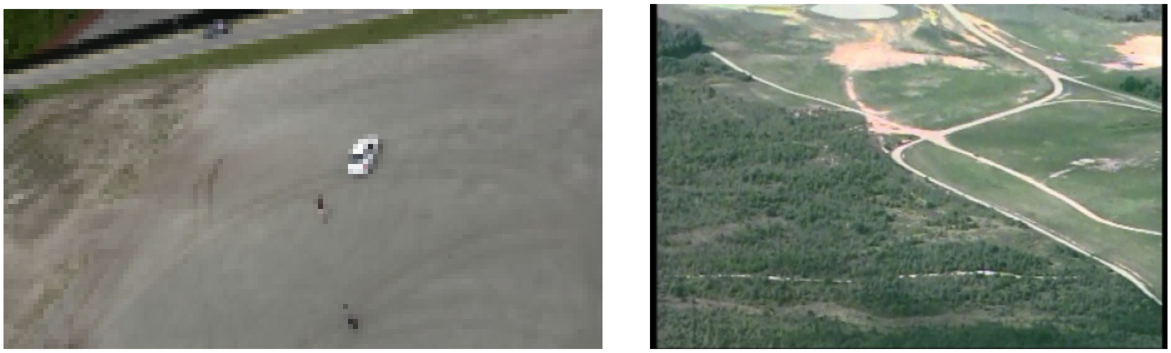

Figure 7-1 Example UCF dataset (left) and Virat dataset (right)

UCF-ARG: http://crcv.ucf.edu/data/UCF-ARG.php

UCF-ARG (University of Central Florida-Aerial camera, Rooftop camera and Ground camera) Data set is a Multiview Human Action data set. UCF-ARG consists of 10 actions performed by 12 actors recorded from a ground camera, a rooftop camera at a height of 100 feet, and an aerial camera mounted onto the payload platform of a 13' Kingfisher Aerostat helium balloon. The 10 actions are Boxing, Carrying, Clapping, Digging, Jogging, Open-Close Trunk, Running, Throwing, Walking and Waving. Except for Open-Close Trunk, all the other actions are performed 4 times by each actor in different directions.