Estimation of vehicle movement in urban areas from thermal video sequences

E. Michaelsen and U. Stilla

Abstract—In this contribution we describe a method to assess the activity of vehicles based on airborne image sequences taken by an infrared camera. At the resolution of approximately one meter vehicles appear as elongated spots. Active vehicles often appear as a configuration of a dark and a bright spot close to each other, where the latter results from the warm motor hood. For the estimation of vehicle movement a precise measurement of the sensor movement is required. The sensor movement is inferred from image sequences. A robust estimation method avoids outliers in the set of point correspondences. A special selection strategy prefers samples that promise high estimation precision.

Index Terms—Infrared surveillance, Machine vision, Object recognition, Urban areas, Vehicles.

I. INTRODUCTION

Vehicle detection has been an important topic in computer vision for a long time [1]. Optical flow helps a lot in segmenting the moving vehicle from the stationary background. Such successful approach is still being pursued, even if a geometric model of the vehicle is utilized [4]. Detecting vehicles with such a model from a single image in cluttered environments spends most effort for the discrimination of vehicle features from clutter features 0. This can be reduced if the detection approach strongly relies on the shadow cast by the vehicle [6]. The advantage of using videos and a simple threshold on the optical flow disappears if not only the vehicle is moving, but also the camera. But a scene fixed camera will only capture the activity in a certain very limited area. An airborne camera is much more flexible and can cover large areas. Vehicle detection from airborne videos has also been addressed [9].

Apart from movement temperature is another important cue to active vehicles. Furthermore, thermal images provide the opportunity to reveal the activity in an urban area by day and night. We propose to use an aircraft with a thermal camera for estimating the vehicle activity in urban terrain. The appearance of vehicles with this sensor depends on many factors, e.g. the daytime and the engine temperature. Passive vehicles often appear as single spots darker than the surroundings. They appear grouped into rows along the margins of roads or in parking lots. A stationary bright spot within a row of passive vehicles can be interpreted as a part of a vehicle that is still warm (has been moved short time before) or as warm spot on the bare concrete giving the hint that there has been a vehicle short time before, that moved away. Active vehicles often appear as a pair of spots. This configuration consists of a bright spot resulting from the engine and a darker spot close to it resulting from the rest of the vehicle. In urban areas other objects may have the same property. But, the evidence for a car will be high, if such a pair of spots is moving along the track of a road in the correct direction.

A. Geocoding

For geocoding of the thermal image we use a projective transformation. We assume that the aircraft is equipped with GPS/INS facilities, so that a good initial estimation for its position and orientation is given. Internal parameters of the camera, like its camera constant, principal point and distortion parameters may be calibrated off-line in advance. Automatic geocoding is based on the same principles as the automatic sensor movement estimation below.

B. Grouping

A single dark or bright spot object on the margin of a road will be a vehicle with certain evidence. We argue that this evidence will be much higher if the spot is member of a row of spots equidistantly placed along the margin. This justifies the use of grouping techniques on the spots. A pair of a bright and a dark spot very close to each other may be interpreted as a single active vehicle. We use production nets for describing these grouping tasks, which are guided by the map to avoid problems with combinatorial search [7] [10].

C. Mobility

Video sequences allow stabilization of the detection of stationary vehicles and estimation of the velocity of moving vehicles. If no map knowledge is available, the sensor movement has to be estimated from the image sequence alone. In urban areas there is usually enough structure in the thermal image sequence to make such estimates feasible. Most vehicles in urban areas move much slower than the aircraft from which the images are taken. Therefore the requirements for the precision of the estimation are high.

E. Michaelsen is with FGAN-FOM Research Institute for Optronics and Pattern Recognition, 76275 Ettlingen, Germany (e-mail: mich@fom.fgan.de).

U. Stilla is with FGAN-FOM Research Institute for Optronics and Pattern Recognition, 76275 Ettlingen, Germany (e-mail: stilla@fom.fgan.de).
II. SENSOR MOVEMENT

A. Interest point detection

The estimation of sensor movement relies on the correspondence of image features from one image to another one. Correspondence of local image features is assessed by correlation. This can be done for immediate successors or for image pairs that are further apart in the sequence. The latter has the advantage, that bigger displacements with the same positioning error will usually yield more precision. If we restrict the features to point-like objects, we should avoid positions where the image is homogenous. Also, because of the aperture problem it is not recommended to use positions, where a 1D-feature (e.g. edge or line) is present. We call all the other positions in the image interest points which are suitable for sensor movement estimation.

Many interest points can not be found by a spot detector. This includes vertices of building outline polygons and T-, Y- or X-crossings of line-shaped features. For such objects we use the interest-operator proposed by Foerstner [3]. This filter is specially tailored to the classification of image positions into the three classes homogenous (0D), oriented (1D) and interest points (2D)(i.e. suitable for stable 2D-correlation results).

Fig. 2b shows the non-zero pixels that the interest operator extracts from the image section displayed in Fig. 2a. Notice that, e.g. the building outline vertices on the left margin of the section appear here, while they are missing in the spot filter results displayed in Fig. 2c and Fig. 2d.

B. Feasibility of Movement Estimation

Sometimes the transformation of the frames of an airborne video sequence to a world coordinate system is not given or it is not given in sufficient precision. In such situations the movement of the sensor with respect to the world has to be estimated. With this information the movement of the vehicle can be inferred from its apparent movement. This is only possible, if the majority of the interest points belongs to static structure in the scene. In urban areas most structures will result from buildings. Fig. 1 shows such a situation. While most of the scene is following a collective movement resulting from the sensor movement, a human subject can easily detect the movement of the car on the horizontal road as outlying.

Measurement of sensor and object movements will fail, if many of the interest points are dynamic in the scene (e.g. sequences of a motorway in open area, where a large portion of interest points will result from moving vehicles). For the estimation of sensor movements from an airborne video we distinguish two cases: spatial appearance and flat appearance.

C. Spatial appearance

If the scene has considerable depth with respect to the height of the sensor and the precisions of the measurements, we will follow the approach of Faugeras et al. [2] and construct fundamental matrices from the correspondences of a set of interest points. Non-linear fundamental matrix estimation needs at least seven interest points correspondences. We use the linear algorithm that needs at least eight correspondences. The fundamental matrix with subsequent linear triangulation provides a complete scene and trajectory reconstruction up to projective 3D-homographies [5]. However, estimation of vehicle movement requires a Euclidean reconstruction. This can only be acquired automatically, if additional knowledge or fusion with data from other sensors is applied.
D. Flat appearance

In the other case (the scene appears flat) the calculation of fundamental matrices becomes unstable. Then an estimation of a projective 2D-homography between the frames is much more appropriate. This estimation can be made on only four point correspondences. It is feasible with interest lines also instead of points.

In general it is possible to detect more than four point correspondence candidates in the data. But not all of these correspondences are correct. One or two outliers already severely distort the resulting estimation which is a severe problem particularly with projective homographies. A proper subset of the set of correspondences has to be searched.

A possibility to avoid this problem would be to identify outliers by sequentially removing or re-weighting suspicious correspondences and testing the resulting estimation for major changes. Some authors recommend the RANSAC method for this purpose instead [5].

The precision of an estimate of a homography is dependent on the position of the point correspondences. These should cover as much of the whole region as possible. A sample with all points in a small region or even nearly aligned in one straight will give large errors. The RANSAC method usually ignores such qualities that may easily measured simultaneously with the calculation of the estimation.

E. Good Sample Consensus

We decided to improve RANSAC by considering the quality of corresponding samples. We call this strategy GSAC (good sample consensus). Estimation votes are calculated from minimal samples containing four correspondences. Each correspondence is assessed according to its correlation value. Good correspondences are preferred for further processing. Pairs of correspondences are constructed and evaluated according to the distance between the points. A larger distance is assessed better than a small distance.

Then quadruples of points correspondences are formed. Each quadruple is assessed according to the area content of the smallest triangle in it. Thus configurations with three collinear points will gain a low assessment. Each quadruple is voting for a projective homography. Good votes with respect to the criteria mentioned start the search for clusters in an accumulator for homography estimations. All of these calculations, evaluations and constructions are implemented in productions. The order in which the productions are used on the objects is given by a data driven control scheme.

Fig. 3 displays the general interaction of productions and objects by a net. Particularly the voting for homographies by GSAC is established at the top of the left production net in production p_{5}. Fig. 4 shows the best evaluated homography estimation on the example image pair from the airborne IR-sequence. It is based on a cluster of 12 estimations close to each other with respect to a special matrix metric. The estimations are based on quadruples of correspondences where the union of all correspondences in the cluster contains 9 correspondences. These are displayed as black lines in Fig 4. The quadruple construction and evaluation knowledge is captured in production p_{4}. Production p_{3} forms the correspondence pairs. The correspondence of interest points from different frames is performed by production p_{2}.

The interest operator is an iconic operation. It sets most of the image to zero and leaves small non-zero islands (see Fig. 2b). These are grouped into spots by production p_{1}. Such spot has only one position attribute, the center of mass, given with sub-pixel accuracy. Fig. 2f displays spot objects formed from the pixel objects shown in Fig. 2b.

The presented objects are not the result of an exhaustive search with the productions of the net. The search is terminated after a pre-determined number of steps. Full exhaustive search is usually not feasible, because of the combinatoric nature of the search.

III. VEHICLE MOVEMENT

A. Spot detection

A spot detector [7] looks for maximal and minimal gray values in rings around bright pixels. If the surrounding pixels in one ring are all significantly darker, then there is evidence for the presence of a bright spot in that location. Fig. 2d shows...
the non-zero pixels resulting with the hot spot filter from the section presented in Fig. 2a. Dark spot pixels presented in Fig. 2c are found in an analogue way.

**B. Marginal Conditions of Vehicle Movement Estimation**

A video sequence of thermal images allows confirming the results of the single image analysis. While the movement of vehicles can be measured directly in such data, features like membership to a group and stable scene position, which have proven useful for the stationary vehicle detection, are not valid for moving vehicles. The search for the position of moving vehicles and the direction of their movement can be significantly reduced by the utilization of the road layer of GIS-data. This localization constraint differs from the constraint used for the positions of stationary vehicles. A stationary vehicle may be parked close to the margin on a road even partially on the sidewalk [7]. A moving vehicle is constrained to the interior region of the road.

**C. Subtracting Global from Apparent Movement**

While the left production net in Fig. 3 estimates homographies, the right production net is constructed to detect vehicles and determine their apparent movement. It works on the spot filter input, the production \( p_6 \) is structurally very similar to production \( p_1 \). Fig. 2e presents the hot and cold spot objects resulting from this production. There are much less spot objects (hot and cold) than interest position objects. The formation of pairs with production \( p_7 \) is different from the formation of correspondence pairs in production \( p_3 \). In production \( p_7 \) the spots have to be very close together and one has to be a hot spot, while the other one has to be cold. In production \( p_3 \) the interest positions where taken preferably far apart. Production \( p_8 \) finally captures the correspondence necessary for the apparent motion estimation. Like in production \( p_2 \) the assessment is based on correlation.

For the absolute movement estimation we subtract the apparent movement of each car cue correspondence object from the current best assessed estimation found applied to that position. A decision whether such a cue is regarded as mobile is based on a threshold on the residual movement. A correspondence cue earlier identified as a mobile may later be rejected based on a different homography estimation with more confidence in it.

**IV. EXPERIMENTS**

The used imagery was taken by an aircraft cruising over a dense urban area in a very flat terrain. The thermal camera was nearly looking to nadir direction. The flight altitude was much higher than the building heights in the scene. Due to the flat terrain we could only test our flat appearance approach.

Fig. 4 shows a result we obtained on a pair of frames 10 time steps apart. The 12 parallel black lines display interest point correspondence objects. These objects make up the best assessed sample which is found after the predefined number of search steps. The projective, rotational and scale components of the resulting homography are small. The most prominent component is a translation of approximately 80 pixels in vertical and 8 pixel in horizontal direction. Only one car cue is inferred from this image pair. It is drawn in Fig. 3 as pair of white lines. The short thick line indicates the orientation of the car. The long thin line indicates its apparent motion. There is hardly any horizontal component in this vector. So there is a residual movement of about 7 pixels consistent with the orientation of the car. From this we infer a speed ten times slower than the aircraft, but still significant enough to classify this as a mobile object. Of course, such an inference is only accepted if it turns out stable over a certain number of frames of the video.

**V. DISCUSSION**

The number of frames in which a moving vehicle is visible depends on its moving direction and velocity. Vehicles moving in flight direction can be tracked for a longer time than vehicles in the opposite direction. Particularly vehicles
appearing only for a small number of frames close to the image margin are difficult to detect. Homography estimates are less precise close to the image margins. The impact of an error in the homography estimate is much more severe close to the image margin or outside of the area covered by the underlying sample. Particularly the assessment of the precision of the measurements influences the reliability of the inferences automatically drawn from the data. This topic needs further investigation.

For the estimation of activity in urban terrain we suggested to use two major properties – temperature and movement – which can be determined automatically from thermal aerial video sequences. In [7] we presented a method to perform perceptual grouping on stationary vehicles along the margin of roads. The properties exploited by the two approaches are mostly independent. The approaches complement each other and can be combined.

REFERENCES


Fig. 4. Result of a simultaneous automatic estimation of sensor (black) and car (white) movement on the example video (image pair with 10 frames difference)