PROBLEMS IN GEOMETRIC MODELLING AND PERCEPTUAL GROUPING OF MAN-MADE OBJECTS IN AERIAL IMAGES

Eckart MICHAELSEN, Uwe STILLA
FGAN-FOM Research Institute for Optronics and Pattern Recognition
D 76275 Ettlingen, Germany
{mich,usti}@fom.fgan.de

Working Group III/4

KEY WORDS: Image understanding, Semantics, Knowledge representation, Object recognition, Urban objects.

ABSTRACT

Structural approaches of pattern recognition are frequently proposed for man-made objects in aerial images. Especially model based methods are considered using geometric, topologic and structural knowledge by means of polyhedral parametric or generic models. We review some aspects of perceptual grouping with respect to these recognition tasks. Emphasis is put on some severe problems encountered in the application. No new results or methods are reported. Instead a qualitative comparison and discussion is given between two such approaches with respect to these problems. Also the importance of associative access techniques is stressed.

1 INTRODUCTION

Research on automatic object recognition in satellite and aerial images has been published for several decades now. Especially multi-spectral pixel classification has reached significant maturity, whereas non-local recognition after decades of research still remains a challenge.

1.1 Model based Approaches

For man-made objects like roads, buildings or vehicles frequently model based approaches have been proposed. Models are described in terms of relations of parts forming hierarchical arrangements. Additionally to the part-of relation other relations such as adjacency, parallelity e. c. are used. The recognition process then consists of corresponding sub-steps of perceptual groupings, searching for constructions consistent with the model.

1.2 Applications

This approach has been proposed for quite different recognition tasks. Detection (i. e. searching for special objects), classification (i. e. assigning object class labels to data objects) and even reconstruction of all objects in a scene have been investigated. Data may have been acquired by different sensors (e. g. visible light, IR, SAR or laser range data). Perspectives vary from very oblique to perpendicular. Objects of interest may be fixed in position (e. g. buildings) as well as moving around (e. g. vehicles or containers). Buildings may have numerous variations in shape whereas vehicles may be categorised into quite homogenous classes. Civil tasks differ frequently from military tasks in the type of objects, sensors and perspectives used. This implies distinct object descriptions and strategies of modelling and searching.

1.3 Difficulties

The euphoric activism of the 80s concerning the application of model based approaches to automatic photogrammetry, remote sensing and cartography has decreased during the last years. This results not from a saturation process settling the research on secure ground and common views. On the contrary, there seems to be resignation because of practical problems with structural models and lack of success in urban scenery. This paper discusses typical problems in the construction and application of such grouping approaches to man-made object recognition in aerial images.

2 GROUPING

To capture our topic more precisely, we have to explain, what is meant by the term ‘grouping’. Two important aspects are the part-of hierarchy and the topological and geometric relations.

2.1 Choosing a Decomposition

If an object of concern may be understood as conglomerate of elementary parts, then we have to consider the space of all possible sub-set decompositions for an appropriate part-of hierarchy. There are very many different strategies in decomposing object aggregates of considerable size into object parts. A simple parallelogram shape may for instance be constructed from a pair of parallel line pairs, an angle pair, a quadrupel of lines e. c.. The choice of the decomposition may have tremendous impact on the search performance resulting. For instance if the decomposition uses parallel line pairs, and the model is used on image data from a rural area with ploughed fields, the search may need to consider many meaningless groupings, and we would have been better off with modelling angles. Usually the choice of the decomposition is left to the applicator. He or she will prefer decompositions that seam natural with the objects under concern. The simple example already showed, that these might not be optimal for the recognition task.

2.2 Perceptual Grouping

In perceptual psychology people have investigated the rules that human vision presumably uses for composing complex objects for nearly hundred years now [Wertheimer 1912]. We refer to the terms of Gestaltist psychologists, when we list the relations used as follows:

Proximity: Spatial neighbourhood is the most important relationship for all non-local pattern recognition. Usually there is at least one parameter accompanied with this relation: The size of the region that is declared to be within the proximity of an object. The choice of a definite value for such a parameter is again left to the applicator, and he or she may again not be aware of the consequences that a somehow reasonable choice with respect to the objects of concern may have on the search effort. Compared to these difficulties the choice of the metric, i. e. the shape of the search region, is of less impact. For the sake of rotational invariance, which seams desirable for many applications, Euclidian metrics are preferred. Sometimes this is traded for performance, when a maximum metric gives much better search performance.

Good Continuation: This property is usually captured geometrically as location of the parts on a curve. It may therefore be sub-classified algebraically into linear, quadratic, cubic e. c. Traditionally pattern recognition handles such relations by Hough-transforms, with the known difficulties of parameter space tessellation. We would like to mention, that if the parameters of a curve are estimated by means of least squares method, they depend on the choice of a subset of objects in the image or scene, and are sensitive to the inclusion of outliers. Such outliers can only be identified after the parameter estimation. Here the search process itself is confronted with a power-set problem.

Similarity: This relation may be viewed as proximity in the attribute space of the objects. For example houses, that fit into a parameterised model of the type ‘simple gabled’, might be viewed as similar, if they have similar height, length, width and roof angle. Also the orientation might be included. Fig 1 shows an example of successful successive exploitation of a combination of the relations continuation and similarity. In such cases also similar inter-house spacing will be demanded, so that the final parametric description of the instance ‘house-row’ contains much less and much more significant information, than the descriptions of the instances ‘house’ in sum.

Fig. 1: Grouping a 3D-House-Row using the relations similarity and linear continuation.
(3D-House-Instances generated from FLAT stereo benchmark)

Symmetry: This relation has drawn remarkably less attention in the pattern recognition community then in psychology. Gestaltists usually found it to be very important in figure-background discrimination. Maybe this lack of interest is just another fear for difficulties, because searching a data set for axis of symmetry is a very non-local task involving correspondence hypothesis, and thus may again demand high computational complexity.

3 PROBLEMS

Transferring such ideas of perceptual grouping as they are presented by psychologists into code that will run in admissible time and with satisfying results on non-trivial data is difficult. So following list of problems is not meant to be complete and the reader is invided to add his or her own items from their own experience. Also the order does not necessarily reflect our assessment on the impact of the problems.

1. **Likelihood of Appearance of Modelled Features:** Usually polyhedrons are used for the geometric aspect of the models and straight line segments are extracted from the image data. There are a couple of questionable assumptions in the match or identification or correspondence between these. Some of the modelled features will appear, but if they do, they do it for different reasons: 1) A figure-background contour will appear, if the reflectivity of the background is different from that of the object, or if there is shadow cast on the background and not on the object. The left hand side of Fig. 2 exemplifies that such contours are frequently missed. 2) Inner contours will appear, if there is a change either in reflectivity or in reflection (mostly due to different surface orientation with respect to the lighting). The former is relatively rare with buildings. Usually a difference in reflectivity is assumed in road recognition – although it is quantitatively unknown. The later will fail for both complete rows in Fig. 2 if the sun is perpendicular to them. Existing geometric models of objects of interest such as buildings or vehicles have often been assembled for different purposes - not for visual recognition. They tend to fail because many of the features modelled do not appear in image (e. g. if free-formed surfaces are modelled by triangles).
2. **Insertion of Unpredictable Objects:** In nearly every aerial image You'll find surprising objects You'd never thought about. Every person modelling a scene will also be able to name several object classes neglected by the model. In Fig. 2 for example also trees, garages, cars e.c. appear, although the model only captures simple gabled houses and roads. Usually the majority of the features will not origin from the objects modelled. Some of the objects not modelled but present in the scene might be similar to parts of the model.
3. **Occlusiuon:** Many Features demanded by a model based recognition algorithm may not appear in the data because they are occluded or partially occluded by other objects. This is extremely difficult with oblique views. But also in example like Fig. 2 partial occlusion is evident. Especially large portions of the road contour are hypothesised instead of measured.
4. **Determination of an Adequate Feature Subset:** From problems 1 and 2 we learned, that usually only a subset of the features are present and we do not know which. Since we need e. g. some 20 features for discrimination from arbitrary background the Model should have some 50 features. The set of all subsets of between 20 and 30 elements of a set of 50 elements is astronomically large. It is not sufficient to specify a threshold for the percentage or absolute number of features, because for instance a small number of 'the right' features will be enough to stabely infer the presence of a modelled object, whereas a larger number of 'unimportant' features will not do.
5. **3D-2D Invariance:** Many properties of geometric models such as angles, measures, topology, are not invariant under perspective projection. Usually one geometric model may either be used for matching with 3D-features gathered by stereo methods or it may be transformed via hidden line or any other rendering into many appearances or aspects, which are used for 2D matches. Then a single 3D model generates possibly hundreds of 2D appearance models. Matching become instable with such many templates. For this reason we prefer working with 3D-models and with 3D-features (like in Fig. 2). Doing this on aerial images demands correctly calibrated, overlapping image sets. Also correspondence errors may lead to wrong 3D-features.
6. **Erroneous Early Decisions:** Often there are alternatives in the correspondence choice between a certain model part and a certain subset of the features. For instance in situations like in Fig. 2 there will often be more then one line that fit – together with some other already identified lines - into the model of a rectangular part of a roof. One might be tempted to only accept the 'best'. But at this stage of analysis this will be a local criterion. Later, when the house-row is established, another line might fit better. Moreover if such early decisions are made on local criteria, the whole outcome of the search depends on the features and models it used in the begin. Often the correct solution will not be found. On the other hand, if every alternative is kept, the computational effort and demand for memory will grow very badly with the size and hierarchical depth of the model. The following problem point clarifies this.
7. **Combinatorial Growth:** Let us again consider the house-row example: Let the probability for missing lines be 0.25 and the probability for the presence of two competing line instantiations be as well 0.25. So each rectangle will in the average have one segment missing and one double. This gives 2 alternatives for each rectangle and little

less than 4 for each roof, because a roof is formed off a pair of rectangles and nearly all four pairings will fit the model. In the worst case – and praxis is not too far away from that – some 4 alternatives for each house in similar positions and parameter settings mean 16 house pairs, 64 triples e. c. In a scene like Fig. 2 this will lead to millions or billions of alternatives that have to be evaluated and compared. Grouping tends to be exponential with its non-locality. Just keep in mind, that it is not even trivial to pick out and unify those alternatives, that actually refer to the same features in the same model roles but put together in a different search sequence. Mathematically with respect to computational complexity two main types of such grouping are to be distinguished - cyclical and cycle free part-of hierarchies. The former is NP and the latter P (see [Michaelsen, 1998] for a syntactic setting of the problem). This does not mean, that cycle free hierarchies pose no problem in computational effort, because in non-trivial cases the algebraic degree of the bounding polynome will be rather high, and there will be intractability in the presence of large data sets. So everybody makes some decision or pruning at some stage of the search. It is also possible to identify very close alternatives as not competing but as co-operative hints for the presence of the same object. Then these should be clustered into a mean representative instead of viewed as competitors. In the example we decided to do this on the 'house level'. But we do not know if that will be wise for all data.

8. **Discriminate Power of geometric Relations:** We already sketched this problem in Sec. 2.1. If a recognition method based of geometric modelling and thus a geometric relation between features is applied to data that the designer has not jet seen, it may fail because the relations may suddenly also hold for many objects whose presence we already stated under problem 2.
9. **Capturing functionality by Geometry:** For many Applications the function of an object is the desired property of interest for its classification e.g. a building may be used for housing people or storing things or administration A vehicle may be used for military or civil purposes. It is questionable how such properties may be recognisable from geometric properties such as adjacency, height and other geometric measurements.
10. **Model Acquisition:** Success and failure of such model-based methods is presumably more dependent on the skill of the person who made the model than on the method or shell used in the recognition process. Who is going to do that in a desired application? Will there be enough trained personal resources? This becomes more urgent with scenes and models getting more complex.

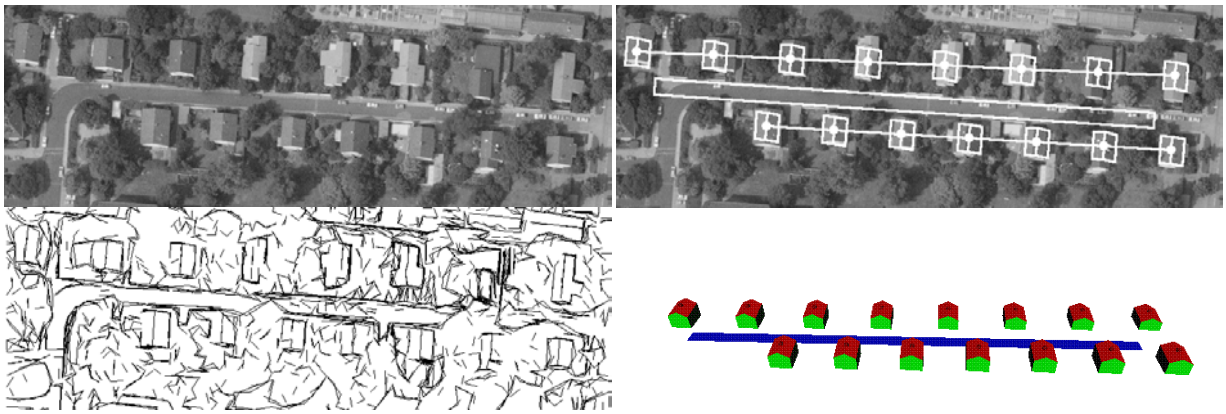

Fig. 2: House Row [Stilla & Michaelsen, 1997]

In Fig. 2 some of the problems mentioned above may exemplarity become evident. Upper left we display a section of an aerial image of a suburban region. Underneath the result of one of our feature extraction processes is shown. Since there is a second image of the same section from another view point (with calibrated geometry) we could generate 3D-lines, angles and structure, finally grouping the house rows and put them in functional connection with the 3D road displayed on the right hand using the principles explained in Sec. 2.

4 COMPARING SHELLS AND SYSTEMS

It is good practice in other pattern recognition disciplines like speech or hand-written character analysis to compare the performance of different methods, approaches and shells on the same benchmark data sets. In the field of concern of this paper, this a laborious task, because not only the input and desired output (ground truth) matters, but also the models and structure used. A comparison then becomes more a qualitative discussion rather than a quantitative competition.

4.1 Some Possible and Published Tools for Structural Model Based Recognition

Many systems have been proposed for model based structural object recognition in the last two decades. Examples are

- The SIGMA system by [Matsuyama, 1985], with productions capturing for instance the functionality between a house, the road and the connecting vehicle path between them. Examples working on single 2D aerial images are given.
- Semantic networks have been proposed to represent the 'knowledge' about the image or scene. Additional to the part-of hierarchy and the geometric relations for grouping then another relation called 'concrete-of' may be defined and help in formulating such things as the different appearance of the same objects in different data sources. One example is the system MOSES [Quint, 1995] capable of combining structure from maps and aerial images of inner city regions. This has been achieved by making use of the shell ERNEST [Kummert, 19??], which is a problem independent semantic network interpreter using A*-search.
- The BPI system has been designed to handle comparably small production systems on huge data sets. In analogy to semantic networks we display such a system as production-net and reported results in 2D and 3D for instance in [Stilla & Michaelsen, 1995]. One speciality is the irrevocable accumulating control, still allowing priority ordered queuing of the search.
- It is also possible to utilise commercial available computer vision shells like KBV with its token sets. Aseptically if there is no need for intelligent control or associative access, an exhaustive search is desired and possible, and a nice interactive working frame helps.
- With rather big amounts of knowledge (production sets) and then inevitably small data sets, also multi purpose AI-interpreters like PROLOG, OPS5, CLIPS, e. c. have to be considered, since they allow to directly give the productions and let the machine do the interpretation on the data.

The choice among these proposals with a given task at hand is by no means trivial. Each has its pro and cons. And understanding any such system takes time and experience. Unfortunately there have been little publications comparing different of these shells or approaches with common benchmarks.

4.2 Comparing Semantic Networks to Production Nets

Since in the recent past we had the opportunity to cooperate and /or compete with F. Quint working on the same data and similar tasks with our BPI that he used for testing his MOSES, we dare to publish some words.

The most serious problem for *production nets* in the recent decade has been problem 1 (missing segments). After all the approach has been syntactically inspired [Michaelsen, 1998]. Parsers will usually reject an input already, if there is any single terminal missing in a syntactically correct pattern. Modelling becomes a laborious task, if also all tolerable kinds of deletions have to be considered. On the other hand this rather explicit handling of problem 4 (defining proper feature subsets for recognition) helps keeping the awareness of its existence. This might be circumvented by using rather flat hierarchies and lots of intermediate cue clustering. The approach then more and more gains the hue of template-matching and Hough-transforms. Of course also problem 7 (combinatorial growth of instance set) is encountered. But the BPI-shell has been specially designed with respect to the handling of large data sets. There are for instance software and hardware mechanisms of associatively querying the database with search regions (see Sec. 4.3 and [Lütjen, Michaelsen & Stilla, 1998]). Also the irrevocable accumulating control scheme helps reducing search effort. Each intermediate result is for instance stored as element instance and may be used by more than one branches of the search. Furthermore production nets allow cycles. Thus they capture generic models like the house-row example in a fairly straightforward syntactic way by some production of the form (house, house-row) – > (house-row).

Semantic networks with search mechanisms like ERNEST seam to suffer most from problem 7 (scaling in computational complexity with the number of alternatives). A search tree is spanned with each node representing a stade of search to be displayed by a partially instantiated network [Quint, 1995]. Every little change in the database like doing a single correspondence between a line segment and a model contour, leads to a new stade. Intermediate results (instances or modified concepts) from one stade are only accessible by the functions working on the successor. If the same intermediate results are needed elsewhere in the tree, they have to be re-established. On the other hand semantic nets are a pretty swift mechanism to handle complicated modelling tasks. Problems 1 and 4 (modelling the appearance and the likelihood of deletions) loose much of their impact compared to how they appeal to the production net designer. Still the awareness of these problems is not hidden. It may even be handled elegantly by the concrete-of-links and other mechanisms like context provided by ERNEST. A serious problem to ERNEST are generic models with cyclic part-of hierarchies. Cycles are explicitly prohibited in the part-of and concrete-of net.

Common to both approaches – but not every other published work on model based recognition – is the awareness of problem 6 (erroneous early decisions, that prune away the correct solution). Both approaches offer means to avoid such difficulties completely, with mathematically proven truth. Since adjusted to this end both approaches suffer seriously from problem 7 (combinatorial growth), both approaches also offer means to relax such demands for strict correctness gradually to the desired degree. Such opportunity to trade problems 6 and 7 against each other is very desirable.

The other problems listed in Sec. 3, namely problems 2 (non-modelled objects), 3 (occlusion), 5 (projection), 8 (adequacy of relations), 9 (visibility of function) and 10 (laborious model acquisition) are more or less problem inherited and independent of the approach chosen.

4.3 Pushing Performance by Inverting Relations into Queries and Search Regions

Most computer systems provide means for performance analysis, so that a statistic on the percentage of run time consumed by each module may be monitored. That will help identifiing administrative overhead. Also a close look on the memory usage helps. A system performing model based perceptual grouping should spend most memory on attribute values of instances and pointers storing interrelations between instances. Most computational efford should be spend on searching groupings and checking the geometric relartions.
We found that our system spends most time on searching partners for a given instance, that will fulfil the relations specified in the productions. Frequently used grouping relations are collinear for line prolongation and for roads and rectilinear for composing pairs or triples for buildings. Of course very important relations are also parallelity, vicinity e. c. The use of inverting techniques (like content addressable retrieval or hashing) for some such relations helps accelerating the search for permissible group partners in the database. Just imagine, that the set of all line instances being located close to a triggering object may be found in time independent of the overall size of the set of all lines. Usually there is a trade-off to be balanced between swift processing and sparing memory [Knuth (1973)]. In structural image analysis today memory is not the problem, so that intelligent retrieval techniques help a lot. Furthermore one might consider solutions including special hardware like [Kohonen (1985)]. We gained some experience in this field in the recent decade [Lütjen, et al. (1998)].
We give again a simple but problematic example, where relation inverting is impractical: Suppose the task is given to group a pair of line objects into a single T-object in 2D image or ground plane space. So the lines should not be parallel and one end of the second line should be close to the other (with Euklidean metrics), like it is exemplified by Fig. 3a). Fig. 3b) then shows the case, that the search queries the base for partners for the bolt drawn upper line triggering. Then only lines with one end in the sketched search region will be able fulfil the relation. That will be quite fiew. So they can be tested sequentially for proper orientation. But if the other line of our group triggers, which is sketched by drawing it bolt in Fig. 3c), we don't have a simple local criterion for the ends of the partner line. Maybe we have to check all lines and thus end up with quadratic complexity with respect to the image size compared to linear like in the case sketched in 3b).

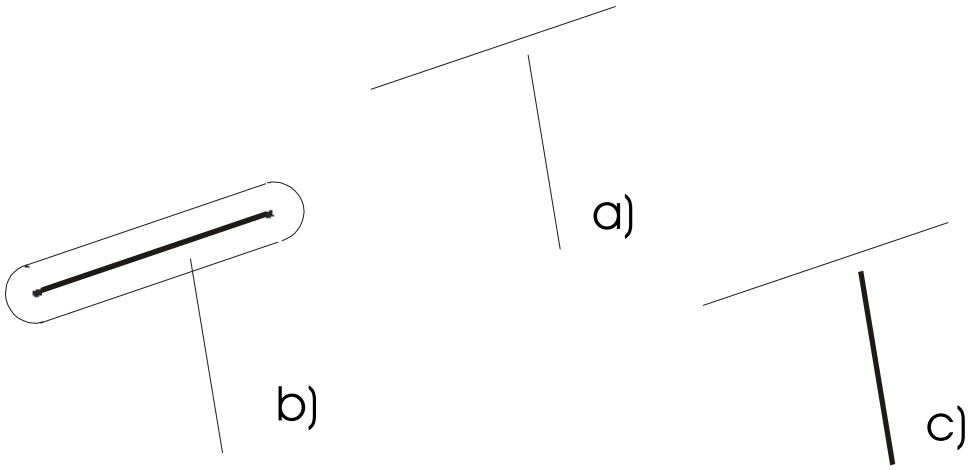

Fig 3: Problems in inverting the geometric relation 'T-shaped'

This simple example shows, that sometimes the sequence in which things are put together matters very much when it comes to complexity assessment. For tasks with very huge data amounts some interdisciplinary fertilisation with data banking is helpful.

5 CONCLUSIONS

We did not present an new approach or method or new results in this paper. We gave a list of ten problems that in our opinion are of importance. With these problems in mind approaches, methods and results should be assessed. We

compared e. g. our system to one other shell and approach, that we know well enough to dare so. The current backlash in the field of structural recognition methods may be explainable to some degree by the difficulties mentioned. But for some tasks, there may be no other way out.
Complex models with many degrees of freedom permit classical one-step template matching (with tolerable effort) only if there is enough prior knowledge (e. g. from maps). Otherwise perceptual grouping of object parts seems to be the only reasonable alternative. This gives best results if an alternation is implemented between grouping and matching. Generic models pose the most severe performance problems because of their inherent combinatorics. We discussed this with an example of grouping buildings and roads into settlement structures. But computational complexity is only one of the difficulties encountered. There are some other serious problems of which the applicator should be aware, if he or she is confronted with a task, that appeals for such a method. Our advice is to check the list of problems with respect to the special task, before one decides to use one of the alternatives mentioned in Sec. 4.1 or something else.
We do not want to discourage anybody. Perceptual grouping and model based recognition remains a fascinating and promising discipline inside computer vision. The more problems you encounter the more ambitious becomes the topic, and awareness of the problems helps to find clever solutions.

REFERNENCES

Ade F (1997) The role of artificial intelligence in the reconstruction of man-made objects from aerial images. In: Gruen A, Baltsavias EP, Henricsson O (eds) Automatic extraction of man-made objects from aerial and space images (II). Basel: Birkhäuser

Knuth DE (1973) The Art of Programming. Vol. 3:Sorting and searching, Reading Mass.: Addison-Wesley

Kohonen T (1985) Content addressable memories. Springer, Berlin.

Kummert F, Niemann H, Prechtel R, Sagerer G (1993) Control and explanation in a signal understanding environment. Signal processing, 3: 111-145

Matsuyama T, Hwang V (1990) SIGMA: A knowledge-based aerial image understanding system. New York: Plenum Press

Metzger W (1975) Gesetze des Sehens: Die Lehre vom Sehen der Formen und Dinge des Raumes und der Bewegung. Frankfurt: Waldemar Kramer

Michaelsen E, Stilla U (1998) Remarks on the Notation of Coordinate Grammars. In: Advances in Pattern Recognition, Joint IAPR Workshops SSPR'98 and SPR'98 in Sydney, Berlin: Springer, 421-428.

Quint F (1996) Recognition of Structured Objects in Monocular Aerial Images Using Context Information. In: Leberl F, Kalliany R, Gruber M (eds) Mapping buildings, roads and other man-made structures from images, IAPR-TC7. Wien: Oldenburg, 213-228

Stilla U (1995) Map-aided structural analysis of aerial images. ISPRS Journal of Photogrammetry and Remote Sensing, 50(4): 3-10

Stilla U, Michaelsen E, Lütjen K (1996) Automatic extraction of buildings from aerial images. In: Leberl F, Kalliany R, Gruber M (eds) Mapping buildings, roads and other man-made structures from images, IAPR-TC7. Wien: Oldenburg, 229-244

Stilla U, Michaelsen E (1997) Semantic modelling of man-made objects by production nets. In: Gruen A, Baltsavias EP, Henricsson O (eds) Automatic extraction of man-made objects from aerial and space images (II). Basel: Birkhaeuser, 43-52

Wertheimer M (1912) Experimental Studies on the seeing of motion. Reprinted in: Shipley T (1961) Classics in psychology. New York: Philosophical Library