Pixels, Stixels, and Objects

David Pfeiffer, Friedrich Erbs, and Uwe Franke

Daimler AG, Research & Development, Sindelfingen, Germany

Abstract. Dense stereo vision has evolved into a powerful foundation for the next generation of intelligent vehicles. The high spatial and temporal resolution allows for robust obstacle detection in complex inner city scenarios, including pedestrian recognition and detection of partially hidden moving objects. Aiming at a vision architecture for efficiently solving an increasing number of vision tasks, the medium-level representation named Stixel World has been developed. This paper shows how this representation forms the foundation for a very efficient, robust and comprehensive understanding of traffic scenes. A recently proposed Stixel computation scheme allows the extraction of multiple objects per image column and generates a segmentation of the input data. The motion of the Stixels is obtained by applying the 6D-Vision principle to track Stixels over time. Subsequently, this allows for an optimal Stixel grouping such that all dynamic objects can be detected easily. Pose and motion of moving Stixel groups are used to initialize more specific object trackers. Moreover, appearance-based object recognition highly benefits from the attention control offered by the Stixel World, both in performance and efficiency.

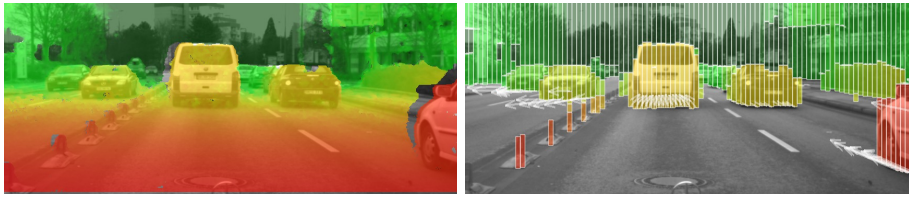

Fig. 1. Dense disparity image and the corresponding Stixel World representing the three-dimensional environment in front of the vehicle. The colors encode the distances, with red = close, green = far away, and gray = freespace. The arrows show the motion vector of the tracked objects. This medium-level representation achieves a reduction of the input data from hundreds of thousands of single depth measurements to a few hundred Stixels only.

1 Introduction

Driver assistance in complex urban environments is a challenging task requiring comprehensive perception of static and moving objects. There is no doubt that stereo vision will play a dominant role in in this context. However, there are two main challenges:

First, the growing number of necessary recognition tasks and their complexity ask for a proper vision architecture that replaces today's independent application-specific modules. This paper proposes such an architecture that bridges the gap between pixels, objects, and the complete representation of the dynamic scene.

Secondly, adverse weather and difficult lightning conditions ask for algorithms of extreme robustness. This can be achieved best if the information contained in the image sequence is optimally exploited. This means that research has to look for algorithms that maximize the probability to find the best solution. The algorithms presented in this paper are a step towards that goal.

Semi-global matching (SGM) [14] outperforms all known local disparity estimation schemes [26]. In [13], Gehrig showed how to run this powerful scheme on energy-efficient FPGA hardware allowing to compute dense high quality depth maps in real-time. On the downside, this high density leads to a large computational burden when the resulting data has to be processed and evaluated by a constantly increasing number of subsequent vision tasks.

In order to minimize the required bandwidth and the computational burden of subsequent vision tasks, we introduced the so called "Stixel World" [2,25], a versatile and extremely compact three-dimensional medium-level representation. As shown in Figure 1, the three-dimensional information is represented by a few hundreds of small rectangular sticks of certain width, height and velocity.

The paper is organized as follows: Section 2 summarizes related work, before the optimization scheme used to compute the (multi-layered) Stixel World is described in Section 3. Section 4 deals with the Stixel tracking for estimating the motion state of individual Stixels. In Section 5, the optimal segmentation of the Stixel World into stationary and moving objects is discussed, before we show in Section 6 that the Stixel World can act as a highly efficient attention control for object classification.

2 Related Work

The most popular method in robotics and driver assistance for representing 3D environments is to project depth information to occupancy grids [7,20] or digital elevation maps (DEM) [17,18]. These maps model the likelihood of the environment to be occupied. They are used to extract scene attributes such as freespace [1] and obstacle information [23], the location of curbs and sidewalks [22,27], and various other scene and application-relevant features.

However, in the light of the quality of today's stereo schemes, even a DEM only represents parts of the information contained in the disparity image. In

order to represent 3D environments more accurately, we introduced the Stixel World that utilizes the fact that most man-made objects can be approximated by planar surfaces with approximately vertical orientation [2]. Position and height of Stixels limiting the free space are computed in a cascade of different algorithms (occupancy grid, freespace, height segmentation) using dynamic programming (DP) [4]. A related procedure has been recently adopted by [5].

By design, these schemes are limited to just extracting the first row of obstacles, this way forbidding the proper handling of partial occlusion, a quite central challenge in driver assistance. Besides that, given the vulnerability of such a cascade of independent processing steps, we have decided to rework the Stixel computation from scratch [25].

The key motivation for this was the work of Gallup et al. [12]. Their objective is to create 3D volumetric object models. Multiple depth maps from different LIDAR scans are accumulated in a single Cartesian histogram-based elevation map. Thereafter, each cell is split into alternating either *"empty"* or *"occupied"* box volumes. By relying on DP, the authors yield the optimal segmentation for every cell of the grid which we consider the major benefit of this procedure.

Felzenszwalb et al. have presented an inspiring scheme [10] as well. They use appearance cues to assign semantic information to image regions. Monocular images are segmented using a continuous upper and lower bound with ordering constraints. The resulting upper part is called *"background"*, the middle region is assigned to *"object"* (e.g. for infrastructure) and the bottom region to *"floor"*. Again, the authors rely on DP and thus guarantee to find the optimum segmentation. However, their approach is limited to a single object per column only.

Further, Liu et al. [19] use appearance cues to assign semantic information to images using a five parts model (top, left, right, bottom, and center, e.g. for segmenting rooms). Their model constraints are very strict and thus inflexible. They rely on a graph cut [6,16] based approach that approximates a 2D optimum.

In [15], Hoiem et al. presented a scheme to assign the labels of type planar (*"ground"*), vertical (*"object"*), and sky to super-pixels. For this purpose, the authors rely on a greedy algorithm while exploiting pairwise patch affinities. They use an appearance-based boosted decision-tree classifier on a trained data set to infer the probabilities for the class affiliation.

3 The Stixel World

Our man-made environments are dominated by either horizontal or vertical planar surfaces. While horizontal surfaces typically correspond to the ground, the vertical ones relate to objects, such as solid infrastructure, pedestrians, or cars.

This perception model is the basic idea of the Stixel World. In the sense of a super-pixel, each Stixel approximates a certain part of an upright oriented object together with its distance and height. In [25], we have presented a probabilistic approach to compute the Stixel World for a stereo image pair in a single global optimization scheme. The problem of Stixel extraction is derived as a classical maximum a-posteriori estimation problem, this way ensuring to obtain the best

segmentation result for the current stereo disparity input. An example result for our method is shown in Figure 1.

Given the left camera image I of a stereo image pair and the corresponding disparity image D (all of size $w \times h \in \mathbb{N}^2$), a multi-layered Stixel World corresponds to a column-wise segmentation $L \in \mathbb{L}$ of I into the classes $C = \{\mathrm{o}, \mathrm{g}\}$ (*"object"* and *"ground/road"*) of the following form

$$\begin{aligned} L &= \{L_u\}\,, \text{ with } 0 \leq u < w \\ L_u &= \{s_n\}\,, \text{ with } 1 \leq n \leq N_u \leq h \\ s_n &= \left\{v_n^b, v_n^t, c_n, f_n(v)\right\}, \text{ with } 0 \leq v_n^b \leq v_n^t < h\ ,\ c_n \in C \end{aligned} \tag{1}$$

The total number of segments s_n for each column u is given by N_u, the image row coordinates v_n^b (base point) and v_n^t (top point) mark the beginning and end of each segment. The term $f_n(v)$ is an arbitrary function providing the depth of that segment at row position v (with $v_n^b \leq v \leq v_n^t$). All segments s_{n-1} and s_n are adjacent. This implicitly guarantees that every image point is assigned to exactly one label.

Modeling all segments as piecewise planar surfaces reduces the function set f_n to a set of linear functions. Object segments are assumed to have a constant disparity while ground segments follow the disparity gradient of the ground surface. The idea of relying on such basic functions is illustrated in Figure 2.

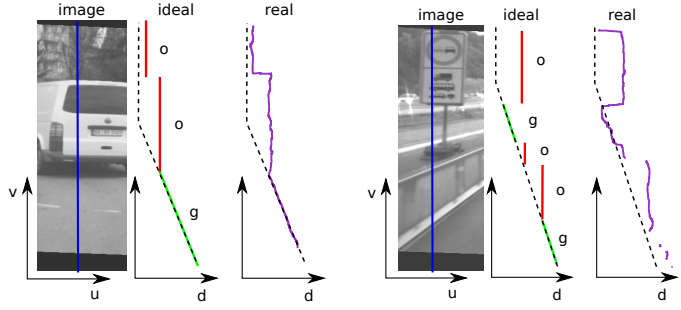

Fig. 2. Data model visualization. The blue line across the image marks the column for segmentation. Red and green denote the ideal segmentation into object and ground. The dashed line is the expected ground profile. The actual disparity measurement vector for the particular scenario is marked using purple.

The computation of the Stixel World is addressed as a MAP estimation problem. Consequently, we seek for the most probable labeling L^*, such that

$$L^* = \underset{L \in \mathbb{L}}{\arg\max}\, P\left(L \mid D\right). \tag{2}$$

Applying Bayes' theorem allows to rewrite the posterior probability $P\left(L \mid D\right)$ as $P\left(L \mid D\right) \sim P\left(D \mid L\right) \cdot P\left(L\right)$. This way, we obtain the product of the conditional probability density of D given L and the prior probability density $P\left(L\right)$.

Here, $P(D \mid L)$ rates the probability of the input D given a certain labeling L and thus represents the data term for the optimization. The second term $P(L)$ embodies the overall probability for each individual labeling L and is the lever to incorporate our world model. In related work, this term is also referred to as the smoothness term. Hence, it allows to support the segmentation with a certain set of physically motivated world assumptions. For example, this includes the following:

- Bayesian information criterion: The number of objects captured along every column is small. Dispensable cuts should be avoided.
- Gravity constraint: Flying objects are unlikely. The ground-adjacent object segment usually stands on the ground surface.
- Ordering constraint: The upper of two adjacent object segments has a greater depth. Reconstructing otherwise (e.g. for traffic lights, signs or trees) is still possible if sufficiently supported by the input data.

To manage the complexity of this segmentation task, neighboring column labelings L_u are considered individually. Further, we treat all disparity measurements $d_{u,v} \in D$ as mutually independent which allows to generalize the disparity input to the vertical disparity vector $D_u \in D$. Additionally, the data input within D_u is stated as independent from all column labels $L_{\hat{u}}$ with $u \neq \hat{u}$. As a result we obtain

$$P(L \mid D) \sim \prod_{u=0}^{w-1} \underbrace{P(D_u \mid L_u)}_{\text{data term}} \cdot \underbrace{P(L_u)}_{\text{prior}}. \tag{3}$$

Finally, dynamic programming [4] is used to infer L^* in real-time. Further details about this Stixel extraction scheme are described in [25].

4 Stixel Tracking

To estimate the motion state of Stixels, the so called "6D-Vision" scheme presented by Franke et al. [11] is used.

In the original form, pixels are tracked over time and image positions as well as disparities are measured at each time step. Assuming a constant motion of the tracked features and a known motion of the observer, a Kalman filter uses these measurements to simultaneously estimate 3D-position and 3D-motion for each tracked feature[1]. Figure 3 shows the obtained result if this algorithm is applied to a turning bicyclist. The arrows point to the expected positions 0.5 sec ahead.

Now, the same principle is applied to precisely estimate the motion state of Stixels [24]. Then, based on this motion information, it is straightforward to determine potential collisions with the ego vehicle (cf. [21]).

Given that in our applications objects of interest are expected to move earthbound, the estimated state vector can be reduced to 4D which is the position and

[1] For more information and illustrative material visit: www.6D-Vision.com

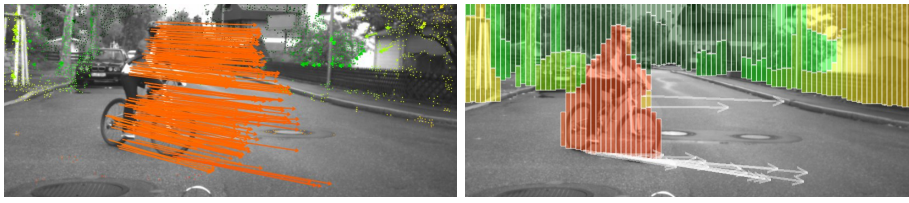

Fig. 3. A bicyclist taking a left turn in front of our vehicle. The left side shows the result when using 6D-Vision point features. The right side shows the corresponding Stixel result.

velocity, i.e. $\underline{X} = (\underline{x}^T, \underline{v}^T)^T$. The dynamic Stixel results for the bicyclist scenario are illustrated in Figure 3 on the right side. As a byproduct of this simplification, the estimation puts fewer requirements on the ego-motion estimation. In practice, the vehicle's inertial sensors are sufficient to compensate for its own motion state. Interestingly, the motion vectors are highly parallel, although the estimation is done independently. This shows the low noise the 6D-Vision scheme achieves thanks to the temporal integration.

5 Stixel Motion Segmentation

Many applications in driver assistance require to know the class type and motion state of other objects. Therefore, in terms of establishing a well structured architecture for modern vision system, gaining this comprehension is an important part of the subsequent processing steps.

So far, Stixels are a compact representation of the local three-dimensional environment, but they do not provide any explicit knowledge about which Stixels belong together and which do not.

With this objective, we have presented an approach that moves towards the grouping of tracked Stixels into the motion classes *"right headed"*, *"left headed"*, *"with us"* and *"oncoming"* as well as *"static background"* [9]. This idea is illustrated in Figure 4 showing the motion segmentation result for the scenario depicted in Figure 1 as well as another exemplary scenario.

Fig. 4. The motion segmentation result obtained using graph cut based optimization. The coloring encodes the individual motion class. Static background is uncolored. The left side shows the result for the scenario depicted in Figure 1, the right shows another traffic situation.

Again, in the style of the described Stixel computation, this task is understood as a MAP estimation problem. Given the dynamic Stixel World, the goal is to find the particular labeling of Stixels into the previously mentioned motion classes that conforms best with our prior knowledge about the current local 3D environment.

Besides assuming rigid motion and striving for spatial smoothness of the segmentation, we use statistics obtained from annotated training data to express where in the scene which type of motion is likely to appear. For this purpose, we model this problem using a conditional Markov random field [6,16] with a maximum clique size of two, thus considering direct neighbor relationships.
The most probable and therefore best class assignment is found by minimizing

$$E = \sum_{i=1}^{N} \Psi\left(l_i^t \mid Z^t, L^{t-1}\right) + \lambda \sum_{(i,j)\in N^2} \Phi\left(l_i^t, l_j^t \mid Z^t, L^{t-1}\right). \tag{4}$$

In this context, $Z^t = \{z_i^t\}$ denotes the Stixel measurement vector at time step t and $L^t = \{l_i^t\}$ are the correspondingly assigned labels. The term Ψ incorporates the unary terms and Φ considers mutual relations between neighboring Stixel labels, such that

$$\Psi\left(l_i^t \mid Z^t, L^{t-1}\right) = \underbrace{p\left(z_i^t \mid l_i^t\right)}_{\text{data term}} \cdot \underbrace{p\left(l_i^{t-1} \mid l_i^t\right)}_{\text{temporal expectation}} \cdot \underbrace{p\left(l_i^t\right)}_{\text{prior}}, \text{ and} \tag{5}$$

$$\Phi\left(l_i^t, l_j^t \mid Z^t, L^{t-1}\right) = \begin{cases} -\log\left(p_{\text{equal}}\left(\Delta_{\text{disp}}\right)\right) & , \text{if } l_i^t = l_j^t \\ -\log\left(1 - p_{\text{equal}}\left(\Delta_{\text{disp}}\right)\right) & \text{otherwise.} \end{cases} \tag{6}$$

The smoothness term $\Phi\left(l_i^t, l_j^t \mid Z^t, L^{t-1}\right)$ is modeled as a Potts model. In doing so, we favor neighboring Stixels to belong to the same label type. The best labeling is extracted using the popular α-expansion graph cut [16]. All control input and algorithmic parameters (e.g. including the weighting factor λ) have been optimized using a large manually annotated sequence data base ($+30,000$ frames) of urban traffic. Further examples for inner city scenarios are illustrated in Figure 5 showing both the tracked Stixel World as well as the motion segmentation results.

6 Object Recognition

A major motivation for the development of the Stixel World was to allow for an efficient attention control of high-level vision tasks. While the easy detection of moving objects as presented in the last section is one example, the recognition of cars, pedestrians and bicyclists is another one which is usually solved by means of classification.

Naturally, the more one knows about the current environment, the less effort must be spent to extract the objects of interest. Accordingly, Enzweiler et al. [8] addressed the exemplary task of vehicle classification (cf. Figure 6) using Stixels and compared this approach against a monocular and a stereo-based attempt.

Fig. 5. Results of the Stixel computation, the Kalman filter-based motion estimation, and the motion segmentation step. Arrows on the base points of the Stixels denote the estimated motion state. The right side shows the corresponding labeling result obtained by graph cut optimization. Furthermore, the color scheme encodes the different motion classes (right headed, left headed, with us, and oncoming). Uncolored regions are detected as static background.

For the monocular approach there are no prior clues about the 3D environment to exploit. Thus, one is forced to test lots of hypotheses spread across the image covering all possible scales that the objects of interest might have. To obtain reasonable results, we used approximately $50,000$ hypotheses per image.

In case of using stereo, the depth data can be easily used to sort out potentially weak hypotheses in advance, e.g. by considering the correlation of depth and scale. This strategy allowed to reduce the hypotheses set by another order of magnitude, such that about $5,000$ remaining hypotheses had to be classified.

Now, using Stixels to control the attention made it possible to address this challenge quite elegantly. Since Stixels inherently encode where in the scene, at which distance, and at which scale objects are to be found, it is just straight-

Fig. 6. The goal is to classify front and rear views of other vehicles. Usually, this requires to spread large amounts of hypotheses across the image. Using Stixels attention classification allows to reduce their amount significantly while, at the same time, the classification performance increases.

forward to directly use this prior knowledge for the hypotheses generation step. This way, we were able to reduce the number of required hypotheses once again by a whole magnitude, such that we ended up at a total of 500 only.

In retrospect, the key conclusion from our experiments is not just that using Stixels allows to speed up the classification by almost a factor of 10 (vs. stereo). It also allows to maintain the detection rate constant, while, at the same time, the number of false alarms decreases by almost one order of magnitude [8]. The potential of this proceeding has also been focus of the work of Benenson et al. [5].

7 Summary and Conclusion

This contribution presented a novel vision architecture for understanding complex traffic situations. Each step from pixels over Stixels to objects is based on global optimization in order to achieve both maximum performance and robustness. In particular, real-time SGM is used to compute dense stereo depth maps from rectified image pairs.

The Stixel world is computed from stereo also in a semi-global optimal manner, only ignoring the lateral dependencies of Stixel columns in order to achieve a real-time implementation. The motion estimation is based on the 6D-Vision principle, taking into account the temporal history of the tracked Stixel.

Finally, in order to extract all moving objects, graph cut based optimization is used to find the optimum segmentation of the dynamic Stixel World. As a result we get an extremely robust recognition system with an unprecedented performance. Practice proves the attempted versatility of the Stixel World. It has been successfully used for the initialization of object-specific trackers like the vehicle tracker proposed by Barth [3].

Muffert et al. [21] were able to easily answer the question whether it is safe to enter a roundabout by analyzing the Stixel World computed for a sideways-looking camera system. As shown in [8], Enzweiler et al. used Stixels to speed up their object classifiers by a factor of ten and simultaneously reduced the false positive rate by nearly one order of magnitude.

References

1. H. Badino, U. Franke, and R. Mester. Free space computation using stochastic occupancy grids and dynamic programming. In *Workshop on Dynamical Vision, ICCV*, Rio de Janeiro, Brazil, October 2007.
2. H. Badino, U. Franke, and D. Pfeiffer. The Stixel World - A compact medium level representation of the 3D-world. In *DAGM*, pages 51–60, Jena, Germany, September 2009.
3. A. Barth, D. Pfeiffer, and U. Franke. Vehicle tracking at urban intersections using dense stereo. In 3^{rd} *Workshop on Behaviour Monitoring and Interpretation, BMI*, pages 47–58, Ghent, Belgium, November 2009.
4. R. Bellman. *Dynamic Programming.* Princeton University Press, 1957.
5. R. Benenson, R. Timofte, and L. Van Gool. Stixels estimation without depth map computation. In *IEEE CVVT:E2M at ICCV*, November 2011.

6. Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *ICCV*, pages 377–384, Kerkyra, Corfu, Greece, September 1999.
7. A. E. Elfes. Sonar-based real-world mapping and navigation. *Journal of Robotics and Automation*, 3(3):249–265, June 1987.
8. M. Enzweiler, M. Hummel, D. Pfeiffer, and U. Franke. Efficient stixel-based object recognition. In *IEEE Intelligent Vehicles Symposium*, Alcalá de Henares, Spain, June 2012.
9. F. Erbs and U. Franke. Stixmentation - probabilistic stixel based traffic scene labeling. In *BMVC*, Guildford, UK, September 2012. BMVA Press.
10. P. F. Felzenszwalb and O. Veksler. Tiered scene labeling with dynamic programming. In *IEEE CVPR*, pages 3097–3104, San Francisco, CA, USA, June 2010.
11. U. Franke, C. Rabe, H. Badino, and S. Gehrig. 6d-vision: Fusion of stereo and motion for robust environment perception. In *DAGM*, Vienna, Austria, September 2005.
12. D. Gallup, M. Pollefeys, and J.-M. Frahm. 3d reconstruction using an n-layer heightmap. In *DAGM*, pages 1–10, Darmstadt, Germany, September 2010.
13. S. Gehrig, F. Eberli, and T. Meyer. A real-time low-power stereo vision engine using semi-global matching. In *ICVS*, Liège, Belgium, October 2009. Springer-Verlag.
14. H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *IEEE CVPR*, pages 807–814, San Diego, CA, USA, June 2005.
15. D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, pages 654–661, 2005.
16. V. Kolmogorov and C. Rother. Minimizing nonsubmodular functions with graph cuts-a review. *IEEE Trans. on PAMI*, 29(7):1274–1279, 2007.
17. I. S. Kweon and T. Kanade. High-resolution terrain map from multiple sensor data. *IEEE Trans. on PAMI*, 14:278–292, 1992.
18. S. Lacroix, I. kyun Jung, and A. Mallet. Digital elevation map building from low altitude stereo imagery. In *Int. SIRS*, 2001.
19. X. Liu, O. Veksler, and J. Samarabandu. Order-preserving moves for graph-cut-based optimization. *IEEE Trans. on PAMI*, 32(7):1182–1196, 2010.
20. H. P. Moravec. Robot spatial perception by stereoscopic vision and 3d evidence grids. Technical Report CMU-RI-TR-96-34, Carnegie Mellon University, 1996.
21. M. Muffert, T. Milbich, D. Pfeiffer, and U. Franke. May I enter the roundabout? a time-to-contact computation based on stereo-vision. In *IEEE IV*, Alcalá de Henares, Spain, June 2012.
22. F. Oniga and S. Nedevschi. Curb detection for driving assistance systems: A cubic spline-based approach. In *IEEE IV*, pages 945–950, Baden-Baden, Germany, June 2011.
23. F. Oniga, S. Nedevschi, M.-M. Meinecke, and T. B. To. Road surface and obstacle detection based on elevation maps from dense stereo. In *IEEE ITSC*, Seattle, WA, USA, September 2007.
24. D. Pfeiffer and U. Franke. Efficient representation of traffic scenes by means of dynamic Stixels. In *IEEE IV*, pages 217–224, San Diego, CA, USA, June 2010.
25. D. Pfeiffer and U. Franke. Towards a global optimal multi-layer Stixel representation of dense 3D data. In *BMVC*, Dundee, Scotland, August 2011. BMVA Press.
26. D. Scharstein and R. Szeliski. Middlebury online stereo evaluation, 2002. http://vision.middlebury.edu/stereo.
27. J. Siegemund, D. Pfeiffer, U. Franke, and W. Förstner. Curb reconstruction using conditional random fields. In *IEEE IV*, pages 203–210, San Diego, CA, USA, June 2010.