Real-Time Event Recognition via Grammatical Parsing

Emmanouil Orfanoudakis and Michail G. Lagoudakis
Telecommunication Systems Institute (TSI)
Technical University of Crete (TUC)
Chania, Crete 73100, Greece

Table of Contents

1 Objective

In the context of the *Noptilus* project, WP6 (Situation Understanding) is defined as:

> Cognitive ability of inferring high-level descriptions and representations of the current state of the environment.

The objective of this work package is summarized as follows:

> Automatically detecting, recognizing, understanding and predicting static underwater features and patterns as well as highly dynamic phenomena and events taking place and influencing the underwater operations.

The project deliverable of this work package is being developed as a system that will serve the following functions:

On-Board On-Line function The on-line system will function as an abstraction level between the *raw sensor data* and the *decision making* module. The goal of the system is to provide:

Robustness against sensor noise and environmental changes in different scenarios.

Abstraction from raw data to facilitate decision making at higher levels.

Compression of sensor information as a result of abstraction.

It should be noted that the description of recognized events has a much shorter representation and, therefore, can be utilized to communicate the state of the AUV in an efficient way over the limited bandwidth of acoustic communication modems.

Off-Board On-Line function Should it becomes necessary, any information on recognized events transmitted by the on-board system (acoustically or through Wi-Fi communication) can be presented to human operators at land stations in ways that enhance awareness about the system and simplifies the supervision of individual vehicles or monitoring of the whole mission itself.

Off-Board Off-Line function The off-line system serves as a training process for generating automatically the configuration of the on-line system (grammatical structure, probabilities, events, etc.). It is intended to serve as the *learning* platform, which will utilize past mission logs, possibly annotated by humans, to automatically derive patterns of normal and abnormal behavior. The goal is to generate a robust event recognition scheme, capable of effective on-line recognition.

2 Definitions

Observation a measurement obtained by a sensor

Observation Stream a temporally-ordered list of observations

Symbol a discrete representation of an observation (or several observations)

(Symbol) Sequence a temporally-ordered list of discrete symbols

Grammar a collection of syntactic rules for producing sequences of symbols

CFG Context-Free Grammar (syntactic rules apply independently of the context)

PCFG Probabilistic Context-Free Grammar (each rule carries a probability value)

Syntactic Tree tree showing how a sequence of symbols is produced by a grammar

Parser software that builds the syntactic tree of a sequence produced by a grammar

Input Sequence sequence of symbols generated from observations fed into a parser

Output Symbol the output of the parser on a given input sequence

Output Sequence a sequence of output symbols delivered by a parser

Grammar Hierarchy output sequences serve as input sequences at higher levels

3 Grammatical Parsing

The power of *Probabilistic Context-Free Grammars (PCFGs)* has been demonstrated in various applications, mainly in the field of Natural Language Processing [1]. Previous work has shown that PCFGs can be utilized effectively [2, 3] for a variety of purposes [4]. Moreover, previous work has demonstrated that for the general problem of *event recognition*, the use of grammars is a viable approach [5–8] and comes with some appealing features:

Expressive Power Most approaches make use of *Markov Models* to express and capture useful events. PCFGs, as a generalization of these models, provide much better ability to describe structure and sequence compared to Markov Models. This property can be illustrated by showing that every event detectable by a (Hidden) Markov Model can also be detected by a PCFG formulation [9].

Compositional Power Given the fact that every parse operation reasons over an input sequence and produces an output sequence of recognized items, it is possible (and desirable) to compose complex, hierarchical recognition schemes, whereby the output sequences of multiple lower-level parsers are fed as input sequence to a higher-level parser, giving rise to PCFG hierarchies.

Robustness The PCFG approach is inherently robust against sensor noise, given the capability of expressing a probabilistic view of all possible input sequences into the rules. Moreover, the parsing operation itself is a *maximum a-posteriori* detector over all possible parses described by the grammar. Therefore, provided that the grammatical and probabilistic description of input scenarios in the PCFG is correct, the output of the parser is an *optimal detector*.

Reusability As long as the environmental assumptions incorporated into the grammar remain valid, a PCFG generated automatically by a learning process utilizing a large amount of past knowledge remains correct and can be reused in future missions for event recognition in similar environments.

3.1 Context-Free Grammars

Context-Free Grammars (CFGs) are formal tools used widely in Computer Science and Language Processing for specifying the syntax of programming and natural languages and for parsing documents for identifying their syntactic correctness and structure. Each CFG is characterized by a set of terminal symbols, which are the base symbols forming sequences, and a set of non-terminal symbols, which are intermediate symbols that characterize structural components of valid sequences. A CFG G is formally defined as a 4-tuple: (V, Σ, R, S), whereby V is a set of non-terminal symbols, Σ is an alphabet of terminal symbols, R is a set of syntactic rules defining how non-terminal symbols can be replaced by sequences of terminal and non-terminal symbols, and S is a start (non-terminal) symbol.

It is possible to construct valid sequences (*strings*) belonging to the language defined by a grammar through a process known as *derivation*. Derivation begins with the start symbol of the grammar and iteratively transforms the current symbol sequence of the derivation by applying rules of the grammar to non-terminal symbols until the sequence contains only terminal symbols. The characterization *context-free* implies that rules can be applied to any non-terminal symbol in a derivation sequence, regardless of the context, that is, the symbols preceding or following the chosen symbol. The power of CFGs is that, given an arbitrary sequence of terminal symbols, a *syntactic tree* can be constructed through a procedure known as *parsing* that describes the structural elements of the input, provided that the input can be derived from the grammar. If that input cannot be derived from the grammar, parsing yields no syntactic tree.

A simple grammar that encodes and produces all sequences of the form[1] $a^n b^n$ or $b^n a^n$

[1] The notation a^n denotes n repetitions of a, for example $a^5 = aaaaa$, whereas a^* implies 0 or more repetitions of a. For a set F, F^* denotes the set of all strings with 0 or more symbols from F.

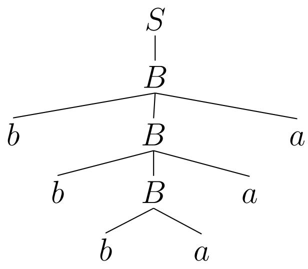
for $n \geq 1$ is the following:

$$
\begin{aligned}
&G = (V, \Sigma, R, S) \\
&V = \{S, A, B\} \\
&\Sigma = \{a, b\} \\
&R = \{S \to A, S \to B, A \to aAb, A \to ab, B \to bBa, B \to ba\}
\end{aligned}
$$

The sequence *bbbaaa* can be derived as follows

$$S \Rightarrow B \Rightarrow bBa \Rightarrow bbBaa \Rightarrow bbbaaa$$

and its syntactic tree is the following



In contrast, the sequence *aabbbb* cannot be derived from this grammar and thus no syntactic tree exists in this case.

3.2 Probabilistic Context-Free Grammars

Probabilistic Context-Free Grammars (PCFGs) extend CFGs by adding a probability value to each rule, so that the values over all rules for a certain non-terminal symbol form a valid probability distribution. During derivation, the choice of the rule for replacing a non-terminal symbol is drawn from the corresponding probability distribution. A PCFG implicitly defines a joint probability distribution over all possible sequences derived by the grammar. This probability value for any given sequence denotes "*how probable it is for the grammar to generate the given sequence*", and therefore how well the grammar *fits* the sequence and vise versa. Sequences that cannot be derived from the grammar have a probability value of 0.

A formal definition of a PCFG G can be given as a 5-tuple:

$$G = (V, \Sigma, R, P, S)$$

whereby:

V is a finite set of *items* (typically represented as alphanumeric strings) that comprise the vocabulary of *non-terminal symbols* of the grammar. They can appear on both sides of a production rule $r \in R$.

Σ is a finite set of *items* that form the set of *terminal symbols* of the grammar, disjoint from V. These symbols form the alphabet for writing sequences derived from the grammar or provided as input to parsing. They can appear only on the right side of a production rule $r \in R$.

R is the set of production rules of the grammar. Formally, R is a finite relation between V and $(V \cup \Sigma)^*$. Each rule takes the form $l \to x$, where $l \in V$ is a non-terminal symbol and $x \in (V \cup \Sigma)^*$ is an ordered set of terminal or non-terminal symbols. Intuitively, each rule allows to *substitute* l with x in any derivation sequence.

P describes the *production probabilities* of each rule $r \in R$, so that the values over all rules with the same non-terminal symbol on their left side form a valid discrete probability distribution.

S denotes the start symbol of the grammar. All valid derivation trees have this symbol as root. Conversely, parsing will have to *reduce* the input sequence to S, when applied to a valid (with respect to the grammar) sequence.

Derivations Every valid (with respect to a grammar) sequence can be generated (derived) by iteratively applying the rules of the grammar to a string w, which is initialized with the start symbol S, until w contains only terminal symbols. During the derivation, every non-empty intermediate string w can be written as:

$$w = \alpha v \beta, \text{ where } \alpha, \beta \in (V \cup \Sigma)^* \text{ and } v \in V$$

A non-terminal symbol $v \in V$ contained in w can be substituted by the right-hand side x of some rule $v \to x$. We say that string $w = \alpha v \beta$ *yields* string $z = \alpha x \beta$ through rule $v \to x$ and we write:

$$w \overset{v \to x}{\Longrightarrow} z$$
$$\alpha v \beta \overset{v \to x}{\Longrightarrow} \alpha x \beta$$

We can generalize this notion, using the closure of the *yields* relation. We say that w yields z if and only if z can be generated from w by applying zero or more rules of the grammar and we write:

$$w \overset{*}{\Rightarrow} z$$

Similarly, if z can be generated from w by applying one or more rules of the grammar, we write:

$$w \overset{+}{\Rightarrow} z$$

Every *recognized* string of a grammar is a sequence $w \in \Sigma^*$ of purely terminal symbols that can be yielded from the starting symbol S:

$$S \overset{*}{\Rightarrow} w$$

A particular derivation tree t of a "$S \overset{*}{\Rightarrow} w$" relation *models* a particular string w.

$$t \models w : \; S \overset{t}{\Longrightarrow} w$$

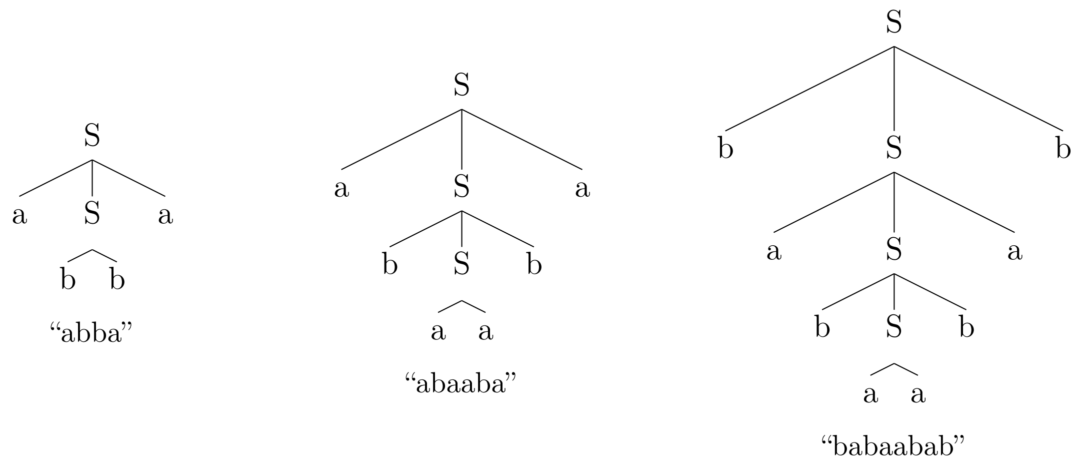


Figure 1: Derivation of various strings from grammar G_p.

Due to grammar ambiguities (see below), there could be multiple trees that model the same string.

A simple grammar, typically used for illustration, is the following:

$$G_p = \big(\{S\}, \{a, b\}, R_p, P_p, S\big)$$

The rules of the grammar R_p (and their corresponding probabilities P_p in brackets) are defined as:

$$S \rightarrow aSa \qquad [p_1] \qquad (1)$$

$$S \rightarrow bSb \qquad [p_2] \qquad (2)$$

$$S \rightarrow bb \qquad [p_3] \qquad (3)$$

$$S \rightarrow aa \qquad [p_4] \qquad (4)$$

whereby $p_1 + p_2 + p_3 + p_4 = 1$. One can easily see that the strings produced and recognized by the grammar G_p are all palindrome strings in $\{a, b\}^*$, such as "abaaba", "abba", and "babaabab". Figure 1 shows the derivation trees for these examples. Strings containing other symbols or strings that are not palindromes cannot be part of the language $L(G)$ of grammar G.

Proper Grammars Not all valid grammars are considered proper. Improper grammars have fallacies in their description, meaning that their formulation is not *minimal*, in the sense of Minimum Description Length. Proper grammars have the following properties:

1. **no inaccessible symbols**: $\forall\ v \in V,\ \exists \alpha, \beta \in (V \cup \Sigma)^* : S \overset{*}{\Rightarrow} \alpha v \beta$.
2. **no unproductive symbols**: $\forall\ v \in V,\ \exists w \in \Sigma^* : v \overset{*}{\Rightarrow} w$.
3. **cycle-free derivations**: $\nexists\ v \in V : v \overset{+}{\Rightarrow} v$.

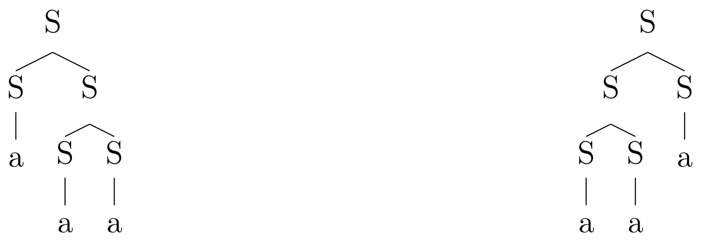


Figure 2: Multiple derivation trees for string “aaa” by grammar G_a.

Inaccessible and unproductive symbols can (and should) be removed from the vocabulary V causing no alteration in the set of input sequences (strings) that a particular grammar derives (or parses). The detection of cycles requires close examination of a grammar and there is no general rule on how to break them. The important subtle detail that arises from cycles in grammars is that, if a cycle exists in the derivation tree of a particular string (or a set of strings), then an infinite number of valid parse trees exist for that particular grammar by recursively introducing the cyclic structure into the parse tree.

Parsing Ambiguities It is possible to construct grammars able to derive a particular string in different ways leading to multiple syntactic trees. Note that the order in which rules are applied during the derivation does not affect the structure of the resulting tree. Therefore, ambiguity is a property with deeper causes in the rules of the grammar. To provide an example, let us define another simple grammar

$$G_a = \big(\{S\}, \{a\}, R_a, P_a, S\big)$$

with the following rules R_a and the corresponding probabilities P_a:

$$S \to SS \qquad [q_1] \qquad (5)$$
$$S \to a \qquad [q_2] \qquad (6)$$

This particular grammar can generate arbitrary sequence of a’s, but there are more than one derivation trees for each sequence. For example, there are two possible parse trees for the string “aaa” that fully comply with the grammar, as shown in Figure 2. Our ultimate goal of event recognition allow us to ignore any difficulties that arise from such ambiguities. When considering the derivation probability of a string (see below), we take into account the probabilities of all possible derivations of this particular string by the grammar.

Sentence and Tree Probabilities PCFGs can be used for modeling stochastic environments in an intuitive and accurate way. Each rule $r \in R_p$ is associated with a probability value $P(r)$. Due to the context-freeness property of the grammars, one can compute the probability of a particular derivation tree for a given string as the product of the probabilities of all the rules used in the derivation. For example, the strings “abaaba”, “abba”, “babaabab” are derived from grammar G_p as shown in Figure 1. The trees are generated from S by applying specific rules:

abba rules 1, 3: $P\big(\text{“}abba\text{”} \mid G_p\big) = P(1)P(3) = p_1 p_3$

abaaba rules 1, 2, 4: $P\big(\text{“}abaaba\text{”} \mid G_p\big) = P(1)P(2)P(4) = p_1 p_2 p_4$

babaabab rules 1, 2 (twice), 4: $P\big(\text{“}babaabab\text{”} \mid G_p\big) = P(2)P(1)P(2)P(4) = p_1 p_2^2 p_4$

Grammar G_p is unambiguous, therefore the trees shown in Figure 1 are unique for each string and the probabilities shown above are the correct derivation probabilities for these strings. For the ambiguous grammar G_a, one can easily see that the probability of each derivation tree of a particular string of length n is:

$$P\left(t \;\middle|\; t \models a^n, G_a\right) = q_1^{n-1} q_2^n$$

Essentially, probability values can be assigned to all trees, whether partial (containing at least one non-terminal symbol) or complete (containing terminal symbols only). The probability of a particular string a^n being generated by grammar G_a is the sum of the probabilities of all its possible derivation trees:

$$P\big(a^n | G_a\big) = \sum_{t:t\models a^n} P\left(t \;\middle|\; t \models a^n, G_a\right) = \sum_{t:t\models a^n} q_1^{n-1} q_2^n = \frac{(2n)!}{(n+1)!n!}\, q_1^{n-1} q_2^n$$

To understand this result, note that, due to our grammar rules, all syntactic trees are binary and the number of binary trees with exactly n leaves is given by the n-th Catalan number [10]. In general, the derivation probability of any given string is computed as the sum of the probabilities of all possible complete derivation trees. Obviously, strings that cannot be parsed by a particular grammar are assigned zero probability, for example $P\big(\text{“}abab\text{”} \mid G_p\big) = 0$.

The derivation probability of a string is useful information in the context of parsing. Given any input string, we can find the probability that this string can be derived by the grammar. If there are multiple derivation trees, we can even find the most probable tree. In general, we can utilize this information to classify several input sequences into different classes based on the probabilities of their respective derivation trees. This property is exploited for the purposes of event recognition.

4 Modeling Events using PCFGs

Utilizing the formal structure of a PCFG, one could create a grammatical structure that can distinguish between different interesting input structures and arrange them according to the probability of occurrence. Let us assume that we need to identify N different events and that a given input sequence w must be *classified* to one of these events. Let us assume a causality relation between the observed sequence and the event that occurred. In other words, we assume that a particular event generates (causes) the particular observed sequence. We shall assume N different grammars G_i, $i = 1, \ldots, N$, each one of which assigns a probability value to any particular input w. Intuitively, if the grammars are indeed *representative* of these N events, then we could simply extract the most probable event E_w that triggered the observation of the input w based on the parse probabilities of each of the grammars. Using a Bayesian approach, we can define the probability of an event E_i given an input sequence w as:

$$P(E_i|w) = \frac{P(w|E_i)P(E_i)}{P(w)}$$

and the extracted event is:

$$E(w) = \arg\max_i P(E_i|w) = \arg\max_i \frac{P(w|E_i)P(E_i)}{P(w)}$$

The probability of the input word $P(w)$ can be ignored in the maximization, since it only serves the normalization of the probability values and remains constant across the different events. Additionally, the probabilities $P(E_i)$ correspond to the *prior probabilities* of the considered events. In the absence of prior knowledge all event can be taken equally probable.

To express the above probabilities in terms of parse probabilities, a relatively weak assumption is made: an event is fully described by the corresponding grammar and no other grammar is correlated to that particular event:

$$P(G_i, E_j) = \begin{cases} 1 & \text{, iff } i = j \\ 0 & \text{, otherwise} \end{cases}$$

Therefore, we can simply substitute E_i with G_i in the decision rule above:

$$E_w = \arg\max_i P(w|G_i)P(G_i) \tag{7}$$

The simplicity of this approach allows an easy interpretation of an input sequence, by simply utilizing the parse probabilities. One can easily verify the following:

- If a particular event cannot generate the observed sequence (or equivalently, the input sequence cannot be derived by the corresponding grammar), then the assigned probability value is zero.
- If two or more events can generate the observed sequence (or equivalently, the input sequence can be derived by two or more grammars), then the most probable event is selected (recognized).

In addition to the above, an important and straight-forward transformation can be used; all event grammars can be combined into a single grammar as follows. Initially, a trivial renaming of the non-terminal symbols of all grammars takes place to ensure that no two grammars have common non-terminal symbols. Then, a single grammar G is constructed by taking the union of each of the elements of the N grammars. A new starting symbol S is created and the set of non-terminals, terminals, and rules is the union of the corresponding sets of the grammars G_i, $i = 1, \ldots, N$. This new grammar is augmented by a set of N rules, which can (must) be applied to expand the new starting symbol to a particular event E_i, thus enforcing the application of the rules of the corresponding grammar G_i through its own starting symbol S_i:

$$S \rightarrow S_i \qquad [P(E_i)] \qquad i = 1, \ldots, N$$

Note that the prior probabilities of the events are incorporated into the new grammar as the probabilities of selecting the rule applied to the root of any syntactic tree, effectively selecting the corresponding event. Given this grammar, a parser must be designed to output which of the N rules was applied at the root of the derivation, thus the one corresponding to the highest parse probability. This rule yields a one-to-one mapping to the N detected events and the parse probability is the event detection probability. The parser may select greedily the most probable parse tree among the possibly multiple N derivation trees.

It is important to note that in the general case, the grammars G_i can be ambiguous. Also, if the grammars describe totally independent sets of words (non-overlapping terminal symbols), then the events can be deterministically distinguished, and this formulation is a bit redundant. If a grammar G_i is the only grammar that can generate (and parse) a particular sequence, then no probabilistic parsing is needed to identify the event, all other grammars cannot parse the string and therefore generate a parse probability equal to zero. We can deduce that all non-trivially *useful* grammars are *ambiguous*, therefore we need a parser that is capable of handling ambiguity.

5 Real-Time Parsing

A parser has been designed and optimized for on-line, real-time operation. Since our grammars are PCFGs, only two families of parsers need to be considered: Earley [11] and Cocke-Younger-Kasami (CYK) [12] parsers. Typically, Earley parsers are considered to be superior to CYK parsers, since for a given input of length n they exhibit an $\mathcal{O}(n)$ complexity. However, for ambiguous grammars, Earley parsers present the same complexity as CYK parsers, $\Theta(n^3)$.

Given the comparable performance of both families, we chose to design a CYK (or chart) parser. Since our observation streams have unconstrained length, only a finite sequence from the stream can be considered for interpretation (parsing) at any time. A parse operation on a sequence of length w (i.e. w consecutive symbols) would yield an event that describes the time interval corresponding to the w symbols in this sequence. Therefore, after an initial mute interval over the first w symbols in the stream, the parser will be able to start recognizing events using the w most recent symbols for any chosen length w. Therefore, our chart parser operates on a *sliding/rolling window* of fixed length w and generates an output (event) to characterize the corresponding time interval. The output sequence of the parser is almost as long as the observation stream (w less symbols at the very beginning).

Our parser exploits the fixed length of the parsing window and the overlap between consecutive windows, namely the fact that at each iteration the oldest symbol is discarded, while the new one is appended to the end of the sequence. Careful bookkeeping is used to store most of the intermediate parse results over a parse window and re-utilize them while parsing the next window. This simple optimization, when applied to the CYK algorithm, reduces the complexity from $\Theta(w^3)$ to $\Theta(w^2)$ at each iteration. Only the parsing of the first window will cost $\Theta(w^3)$, since there is no previous overlapping window. Therefore, in the limit of an infinitely-long observation stream the amortized complexity of each iteration is $\Theta(w^2)$ for some parsing window of length w. Since the rolling window has a fixed length, the complexity of the parser is constant and independent of the length of the observation stream.

Based on the grammar structure model described above, the output of the parser is essentially a symbol indicating the rule, and thus the corresponding event, that needs to be applied to the starting symbol of the grammar for generating the parse tree that maximizes the derivation probability of the input. Preliminary results indicate that indeed the proposed parser can be implemented for real-time operation on embedded systems, such as the Noptilus AUVs.

6 Hierarchical Processing

The output of a parser is a sequence of discrete symbols. Therefore, a straightforward extension of a parsing process is to use its output sequence as input sequence in another, higher-level, parsing process. Even better, several output sequences derived from parsing independent input streams could be combined to form a single input sequence for a higher-level parser, thus giving rise to hierarchies of grammars and/or parsers. Consider, for example, two distinct parsers (built from two different grammars) that operate independently on the observation streams generated by two different sensors. An interesting question that immediately arises is how one could use a third, higher-level grammar to generate a combined interpretation of the two sensor streams by parsing the outputs of the independent parsers. Given the discrete nature of the output of the two parsers, it is trivially easy to generate a new symbol stream, whereby the symbols are all possible combinations of outputs from the two parsers. This stream essentially contains processed, abstracted information from both sensors and, therefore, the third grammar operating on this stream could seek to recognize *higher-level* abstract phenomena or events related to the state of the system, based on the information from both sensors.

Let us denote the possible generated events from the lower-level parsers as $\mathbf{E}^1 = \{E_1^1, E_2^1, \ldots, E_M^1\}$ for the first parser (M events) and $\mathbf{E}^2 = \{E_1^2, E_2^2, \ldots, E_L^2\}$ for the second one (L events). A grammar residing on top of these two parsers would accept as input symbols the Cartesian product of these two sets:

$$\Sigma = \mathbf{E}^1 \times \mathbf{E}^2 = \{E_1^1 E_1^2, \ldots, E_M^1 E_1^2, E_1^1 E_2^2, \ldots, E_M^1 E_L^2\}$$

This modular approach, can accommodate any grammatical parsing scheme, and any event can be "described" structurally using either sensor data or the output of another grammar.

7 Proof-of-Concept Example

To verify and demonstrate the performance of this set-up, an off-line, off-board simulation of the Event Recognition system was implemented in MATLAB, utilizing real sensor data (captured directly by the vehicle during a mission) for generating the observation streams. The target of this example is to recognize the following categories of events:

Diving Behavior These events derive from the depth sensor stream. The goal is to come up with a grammar/parser that describes the motion of the vehicle with respect to depth only. The output could be used for assessing the navigation performance of the platform with respect to diving.

Vertical Orientation These events derive from the Euler angle (pitch) sensor of the vehicle. The goal is to find a grammar/parser that yields a high-level description of the state of the vehicle in terms of pitch. The output could be used for assessing the vertical orientation of the vehicle. is highly correlated to the depth changes of the vehicle, as the propulsion of the vehicle (propeller) is cannot be directed. The pitch of the vehicle must be adjusted (raised or lowered respectively) for the vehicle to ascend or descend.

Vertical Collision These events derive from the combination of the depth and pitch observation streams, seen at the abstract level of the outputs of the two independent parsers discussed above. Note that depth and pitch are highly correlated, since the propulsion of the vehicle (propeller) is not directed and any diving actions (ascend or descend) will have to raise or lower the pitch. The goal here is to come up with a higher-level grammar that parses the combination of the two input streams. The output could serve the detection of faults, such as collisions, in the vertical (downward) motion of the vehicle.

The three grammars/parsers, described in detail below, were tested on a particular logged mission aiming to eventually recognize a known fault event contained in this log: a collision of the vehicle with the sea bottom.

7.1 Preprocessing

For the purposes of our work, we assume that both sensors operate in a roughly synchronous fashion. Indeed, although the two sensors generate values at different rates, the ratio of these rates is fixed: five pitch values are generated by the *inertial measurement unit* in the same time period that one depth value is generated by the *sonar*. A preprocessing step is applied to the raw continuous measurements to generate discrete symbols (observations) as inputs to the parser.

For the depth data, since there is unlimited range, we chose to extract the rate of change instead of the actual depth. To this end, a simple differentiator with respect to time estimates the rate of change at each time using the previous value and the time interval between time steps. In particular, data are preprocessed as follows:

- A total of s raw measurements are collected.
- For each one of them the differentiator generates a rate of change.
- The s rates are averaged and a single rate is produced.
- This rate is quantized: negative (-1), (approximately) zero (0), positive ($+1$).
- The discrete symbol is fed to the parser that processes depth.

For the pitch data, since there is limited range $[-\pi, +\pi]$, we chose to simply quantize its value. In particular, data are preprocessed as follows:

- A total of $5s$ raw measurements are collected.
- The s values are averaged and a single pitch value is produced.
- This value is quantized: negative (-1), close to zero (0), positive ($+1$).
- The discrete symbol is fed to the parser that processes pitch.

This preprocessing procedure generates a sub-sampled stream of discrete symbols for each of the two sensors. Since the sub-sampling is synchronous, both parsers generate output symbols synchronously at a macroscopic time scale. Some degree of jitter is obviously expected to be present in the real-time system, but the variation of the rate of the sensors is in the sub-second time scale and is effectively smoothed out by preprocessing. The size s of the sampling window can be adjusted to meet the required frequency of symbol generation. The higher-level grammar operates on top of these two output streams, therefore there is no need for any kind of data preprocessing.

7.2 Grammars

All three grammars in this preliminary study were designed and tuned by hand, using our intuition about the target domain. Our future plan is to employ an off-line learning method for automatically constructing grammars from data. Such a learning scheme will eliminate the human factor from the process in the future.

The depth parser, using the grammar for depth, outputs one of the following events:

Up The vehicle is ascending (non-terminal symbol U).

Down The vehicle is descending (non-terminal symbol D).

Hover The vehicle is at a roughly constant depth (non-terminal symbol H).

Changing The vehicle is performing a vertically convex or concave maneuver, (C).

Likewise, the pitch parser, using the grammar for pitch, outputs the following events:

Up The vehicle is facing upwards (non-terminal symbol U).

Down The vehicle is pointing downwards (non-terminal symbol D).

Hover The vehicle is roughly level (non-terminal symbol H).

The depth and pitch grammars are defined on the set of terminal symbols generated during the preprocessing step: $\Sigma = \{-1, 0, +1\}$. The starting symbol is S and the rules of both grammars are shown in Figure 3. Note that the two grammars are independent, therefore there is no problem in having common names for some of the symbols. As explained in Section 4, each grammar is a combination of different parse structures, each capable of recognizing a distinct event. These distinct events can be seen at the first rules of each grammar, which describe the derivation of each kind of event from the starting symbol. These rules are of the form:

$$S \rightarrow E_i \quad [P(E_i)]$$

A sampling window of size s and a parsing window of size w translate to one output event being extracted from the last sw data values of the sensor. The value sw is a design parameter and various combinations were tested to study its effect on the recognition process.

On top of the two independent grammars for depth and pitch, a third higher-level grammar is designed to detect events related to both depth and pitch. This higher-level grammar is defined on an input symbol set which is the Cartesian product of the outputs of the lower-level grammars, whose parsers operate synchronously to each other. The input symbols (terminals) of this higher-level grammar are formed using concatenation of the output symbols with the depth parser output contributing the first (left) symbol and the pitch parser output contributing the second (right) symbol. Therefore, the terminal symbols of the higher-level grammar are:

$$\Sigma_{DP} = \{UU, UD, UH, DU, DD, DH, CU, CD, CH, HU, HD, HH\}$$

Each symbol in this set denotes the events detected by the two parsers at the same time. The rules of the higher-level grammar are shown in Figure 4 and the events detected by the corresponding parser are the following:

OK Normal operation, no interesting event (non-terminal OK).

Collision Collision with the sea bottom (non-terminal Co).

Up Draft An upward current pushes the vehicle up (non-terminal UpD).

Down Draft A downward current pushes the vehicle down (non-terminal $DownD$).

$S \rightarrow$	C (**Event C**)	[0.1]
$S \rightarrow$	U (**Event U**)	[0.3]
$S \rightarrow$	D (**Event D**)	[0.3]
$S \rightarrow$	H (**Event H**)	[0.1]
$D \rightarrow$	Dd	[0.5]
$D \rightarrow$	d	[0.5]
$U \rightarrow$	Uu	[0.5]
$U \rightarrow$	u	[0.5]
$C \rightarrow$	UHD	[0.5]
$C \rightarrow$	DHU	[0.5]
$H \rightarrow$	Hh	[0.34]
$H \rightarrow$	h	[0.34]
$H \rightarrow$	DHU	[0.16]
$H \rightarrow$	UHD	[0.16]
$d \rightarrow$	$+1$	[0.98]
$d \rightarrow$	0	[0.02]
$u \rightarrow$	-1	[0.98]
$u \rightarrow$	0	[0.02]
$h \rightarrow$	-1	[0.02]
$h \rightarrow$	$+1$	[0.02]
$h \rightarrow$	0	[0.98]

(a) Depth

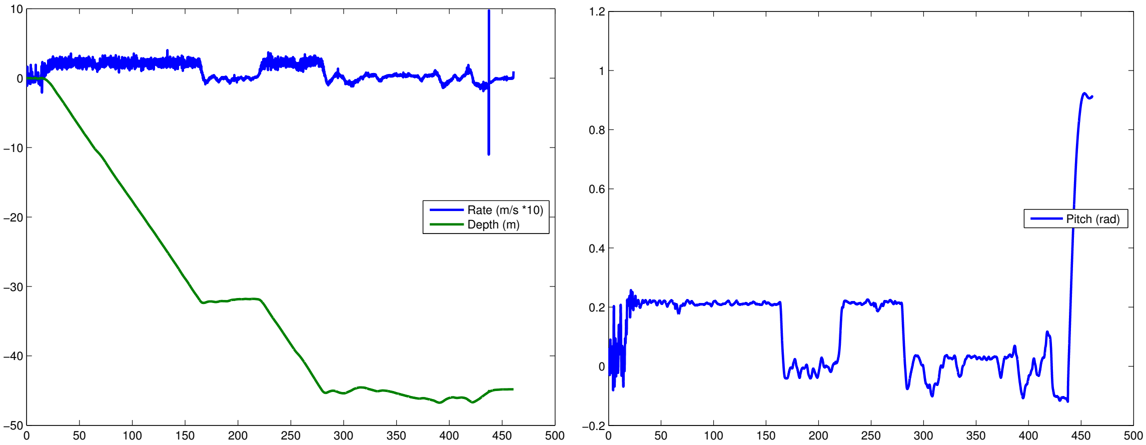
$S \rightarrow$	U (**Event U**)	[0.3]
$S \rightarrow$	D (**Event D**)	[0.3]
$S \rightarrow$	H (**Event H**)	[0.3]
$D \rightarrow$	Dd	[0.5]
$D \rightarrow$	d	[0.5]
$U \rightarrow$	Uu	[0.5]
$U \rightarrow$	u	[0.5]
$H \rightarrow$	Hh	[0.34]
$H \rightarrow$	h	[0.34]
$H \rightarrow$	DHU	[0.16]
$H \rightarrow$	UHD	[0.16]
$d \rightarrow$	$+1$	[0.98]
$d \rightarrow$	-1	[0.01]
$d \rightarrow$	0	[0.01]
$u \rightarrow$	-1	[0.98]
$u \rightarrow$	$+1$	[0.01]
$u \rightarrow$	0	[0.01]
$h \rightarrow$	-1	[0.01]
$h \rightarrow$	$+1$	[0.01]
$h \rightarrow$	0	[0.98]

(b) Pitch

Figure 3: Grammar rules for (a) depth and (b) pitch events.

$S \rightarrow$	OK (**Event OK**)	[0.889]	$UpD \rightarrow$	$UpD\ upd$	[0.5]
$S \rightarrow$	Co (**Event Co**)	[0.001]	$UpD \rightarrow$	upd	[0.5]
$S \rightarrow$	UpD (**Event UpD**)	[0.05]			
$S \rightarrow$	$DownD$ (**Event DownD**)	[0.05]	$upd \rightarrow$	$updt$	[0.9]
			$upd \rightarrow$	h	[0.1]
$OK \rightarrow$	$OK\ o$	[0.5]			
$OK \rightarrow$	o	[0.5]	$updt \rightarrow$	HD	[0.5]
			$updt \rightarrow$	UH	[0.5]
$o \rightarrow$	u	[0.3]			
$o \rightarrow$	d	[0.3]	$DownD \rightarrow$	$DownD\ downd$	[0.5]
$o \rightarrow$	h	[0.3]	$DownD \rightarrow$	$downd$	[0.5]
$o \rightarrow$	$updt$	[0.025]			
$o \rightarrow$	$downdt$	[0.025]	$downd \rightarrow$	$downd\ downdt$	[0.9]
$o \rightarrow$	dot	[0.05]	$downd \rightarrow$	h	[0.1]
$u \rightarrow$	UU	[0.9]	$downdt \rightarrow$	$HU\ downdt$	[0.5]
$u \rightarrow$	LU	[0.1]	$downdt \rightarrow$	DH	[0.5]
$d \rightarrow$	DD	[0.9]	$Co \rightarrow$	$Co\ co$	[0.5]
$d \rightarrow$	LD	[0.1]	$Co \rightarrow$	co	[0.5]
$h \rightarrow$	HH	[0.88]	$co \rightarrow$	o	[0.1]
$h \rightarrow$	LH	[0.07]	$co \rightarrow$	cot	[0.9]
$h \rightarrow$	DU	[0.05]			
			$cot \rightarrow$	UD	[0.99]

Figure 4: Grammar rules of the higher-level grammar for depth/pitch events.



(a) Depth Measurements and Calculated Rate

(b) Pitch Measurements

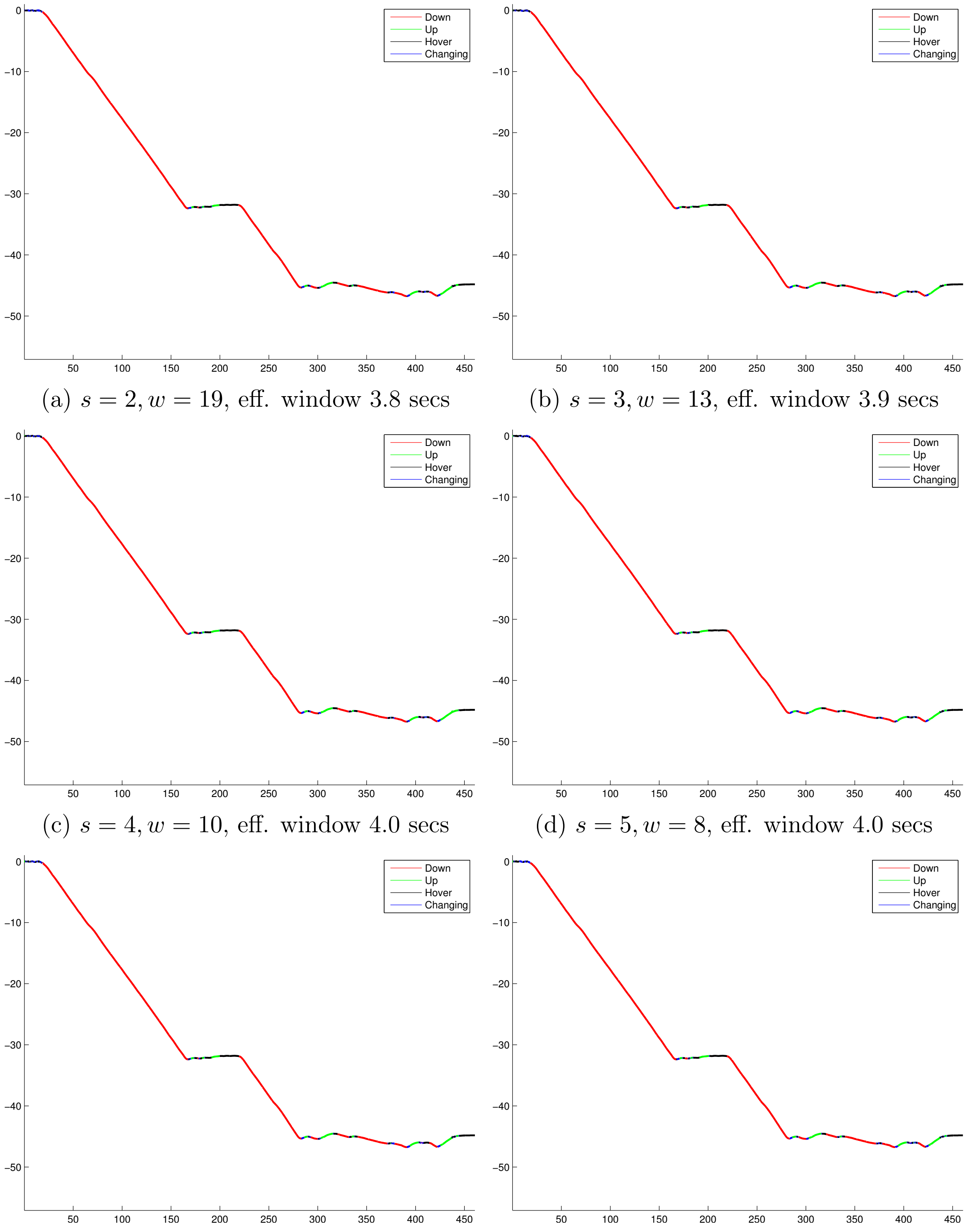
Figure 5: Depth and pitch sensor measurements from the mission log

7.3 Results

The grammars we have described were tested on data from a particular logged mission to verify that we can detect a confirmed fault event on the vehicle: a collision of the vehicle with the sea bottom which occurred at around 450 seconds into the mission. The measurement values of the depth and pitch sensors are graphed over time in Figure 5. The depth measurements are taken at a rate of 10Hz, whereas pitch measurements are taken at a rate of 50Hz. The rate of change in depth estimated by our differentiator (Section 7.1) is also graphed in Figure 5 (a). Our goal is to annotate the mission with useful information about the occurred events. This particular mission is characterized qualitatively by the following:

- The first 10-30 seconds of the log indicate a rough agitated surface. The vehicle moves rapidly around and the pitch sensor is quite noisy at this period.
- The first descend phase (30-160 seconds) is followed by a hovering phase (160-220 seconds) and then by a second descend phase (220-280 seconds).
- After second 280 the vehicle alternates between hover and descend.
- The collision occurs at second 450: there is an evident “bounce" at the depth readings, and the pitch of the vehicle turns rapidly and unnaturally positive.
- After the collision the vehicle aborted the mission and surfaced.

We run our parsers on these data for various values for the sampling window s and the parsing window w. The refresh interval is s times the period of the depth sensor (0.1 seconds) and the effective time window over which event are detected is sw times the period of the depth sensor. For various indicative values of s and w the event detected at each time step (annotated in color) for each of the two sensor data streams is presented in Figure 6 and Figure 7 respectively. In each case, the low-level grammar along with the corresponding parser were used. It is apparent that



(a) $s = 2, w = 19$, eff. window 3.8 secs

(b) $s = 3, w = 13$, eff. window 3.9 secs

(c) $s = 4, w = 10$, eff. window 4.0 secs

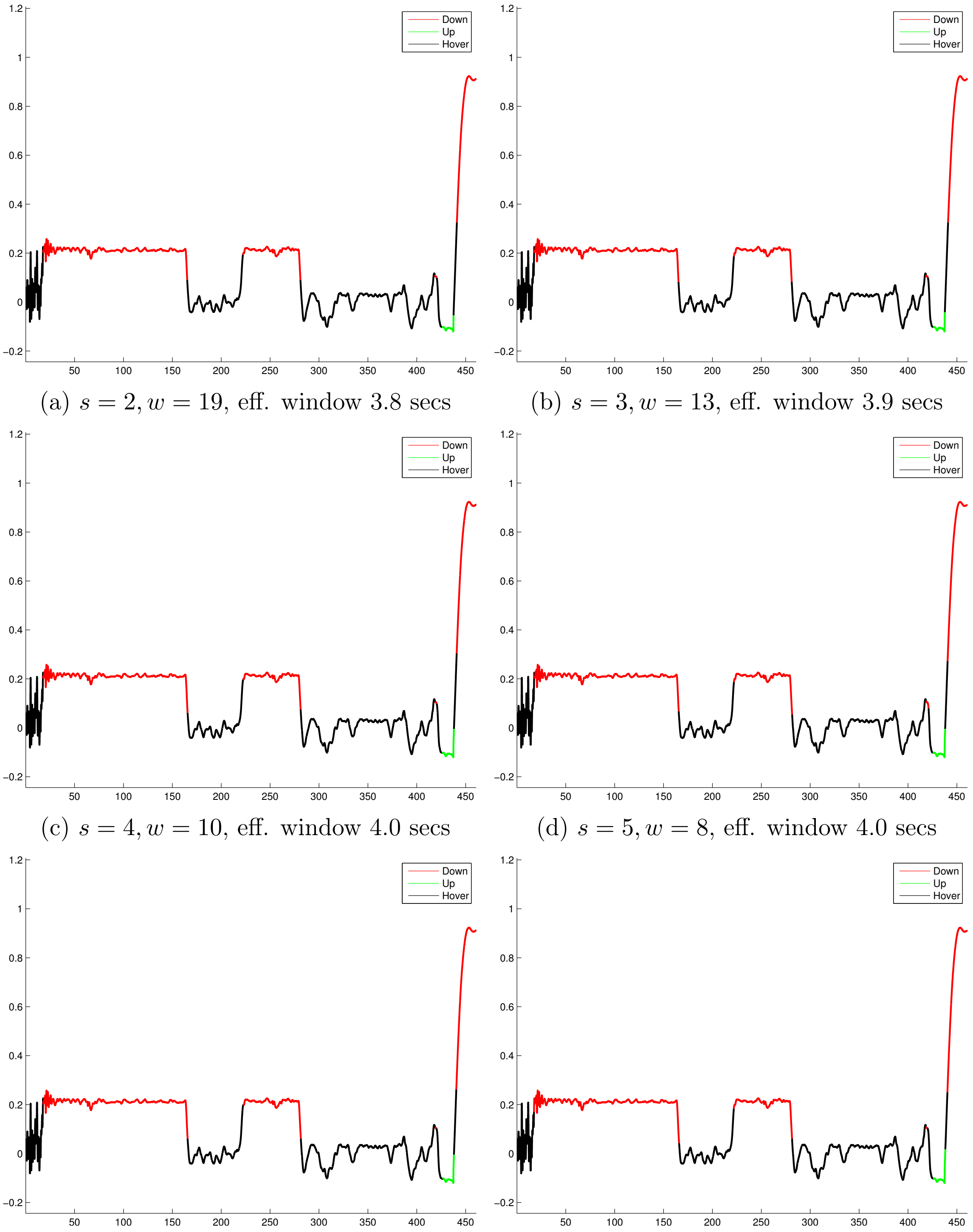
(d) $s = 5, w = 8$, eff. window 4.0 secs

(e) $s = 6, w = 7$, eff. window 4.2 secs

(f) $s = 7, w = 6$, eff. window 4.2 secs

Figure 6: Event annotation of the depth data stream for various time windows.

the parsers correctly annotate the mission data and this annotation is insensitive to the precise choices of s and w, as long as the effective time window remains constant.



(a) $s = 2, w = 19$, eff. window 3.8 secs

(b) $s = 3, w = 13$, eff. window 3.9 secs

(c) $s = 4, w = 10$, eff. window 4.0 secs

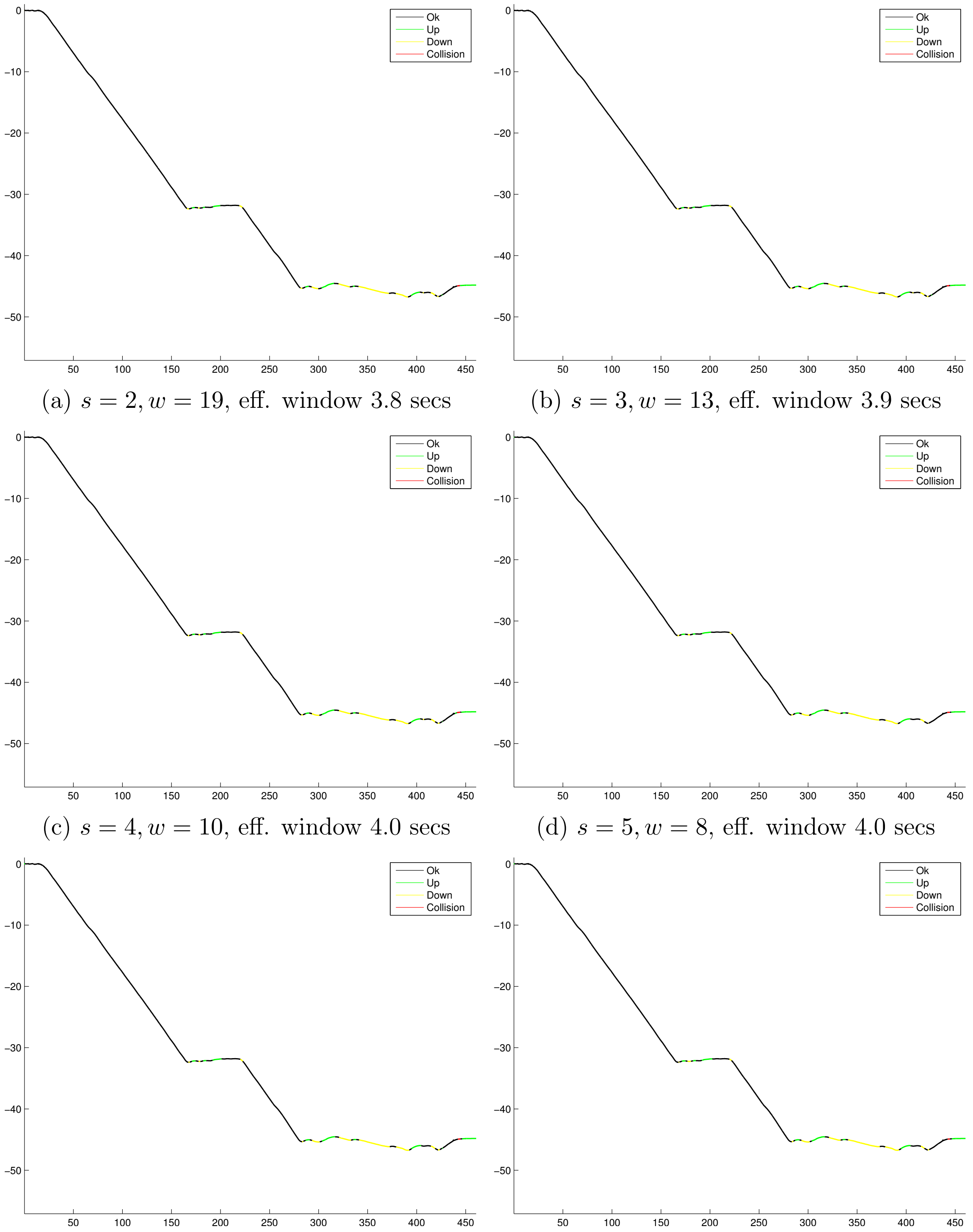
(d) $s = 5, w = 8$, eff. window 4.0 secs

(e) $s = 6, w = 7$, eff. window 4.2 secs

(f) $s = 7, w = 6$, eff. window 4.2 secs

Figure 7: Event annotation of the pitch data stream for various time windows.

The event output of the higher-level parser for various indicative values of s and w is shown in Figure 8, whereby the depth data are now color-annotated by the high-level events. It is worth noting the accuracy of the detected events and especially



(a) $s = 2, w = 19$, eff. window 3.8 secs

(b) $s = 3, w = 13$, eff. window 3.9 secs

(c) $s = 4, w = 10$, eff. window 4.0 secs

(d) $s = 5, w = 8$, eff. window 4.0 secs

(e) $s = 6, w = 7$, eff. window 4.2 secs

(f) $s = 7, w = 6$, eff. window 4.2 secs

Figure 8: Event annotation of the depth data stream using the higher-level grammar.

the detection of the collision at around 450 seconds (shown in red in Figure 8). Once again the results are insensitive to the precise choices of s and w, as long as the effective time window remains constant.

7.4 Performance

The results obtained on this log were quite promising. The annotation is robust to noise and to the selection of the parameters s and w. The desired time abstraction and granularity can be achieved by selecting these two values appropriately, however it is an open question what the right level of time abstraction is. The answer is certainly dictated by the physical aspects of the system and the time extent of the events that need to be recognized. In any case, our approach can accommodate any set of choices and can be applied repeatedly to identify the right level.

The grammars annotate the data in an intuitively correct way. The two lower-level grammars generate a clean and robust abstraction from the raw data. This abstraction should prove indispensable in the long run. Despite the fuzziness of the pitch sensor data, the grammar manages to filter out the rapid motion of the vehicle and infer a hovering event.

One important omission of this toy-example is the positive buoyancy and inertia of the vehicle: the pitch value of the vehicle is out-of-phase with respect to the depth rate of change. This results in miss-detection of upward/downward water draft. This does not relate to the performance of the proposed scheme: it is an omission in the manually-constructed rules of the higher-level grammar. Such phenomena will be accounted for by the automated learning process. Despite this, the verified collision which occurred at around $t = 450$ seconds of the mission is *never missed.*

The performance, simplicity, and scalability of the hierarchical scheme are apparent. The real-time performance of the current parsers (implemented in MATLAB) appears to be feasible. Although the processing time cannot be expected to be trivial, we estimate that the gain in robustness, compression, and modularity will justify the allocation of the some memory and CPU time to event recognition.

In summary, the following qualities of the proposed system should prove invaluable:

Compression of information The detected events can be stored (and transmitted) as an dense and robust abstraction of raw sensor data. As a result, decision making will be greatly facilitated by operating on top of event recognition.

Re-usability and scalability The grammatical description of the sensors can be expanded and reused from mission to mission, generating a library of compact descriptions of both normal and abnormal behavior of various components.

Robustness Although this feat should be further examined and verified, it is clear that our approach is capable of handling multiple scenarios. Grammars can be trained using mission logs and any knowledge acquired can be incorporated into the existing grammatical description.

Powerful representation The grammatical nature of our approach provides intuitive means to express a large number of event structures, sequences, and scenarios.

References

[1] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[2] S. Petrov and D. Klein, "Learning and inference for hierarchically split PCFGs," in *Proceedings of the 22nd National Conference on Artificial intelligence (AAAI)*, 2007, pp. 1663–1666. [Online]. Available: http://dl.acm.org/citation.cfm?id=1619797.1619916

[3] A. Belz, "PCFG learning by nonterminal partition search," in *Proceedings of the 6th International Colloquium on Grammatical Inference (ICGI): Algorithms and Applications*, 2002, pp. 14–27. [Online]. Available: http://dx.doi.org/10.1007/3-540-45790-9_2

[4] G. Neu and C. Szepesvári, "Training parsers by inverse reinforcement learning," *Machine Learning*, vol. 77, pp. 303–337, 2009. [Online]. Available: http://dx.doi.org/10.1007/s10994-009-5110-1

[5] G. Guerra-Filho, C. Fermuller, and Y. Aloimonos, "Discovering a language for human activity," in *Proceedings of the AAAI Fall Symposium on Anticipatory Cognitive Embodied Systems*, 2005, pp. 70–77. [Online]. Available: http://ranger.uta.edu/~guerra/Guerra-FilhoFS05discovering.pdf

[6] D. Lymberopoulos, A. S. Ogale, A. Savvides, and Y. Aloimonos, "A sensory grammar for inferring behaviors in sensor networks," in *Proceedings of the 5th International Conference on Information Processing in Sensor Networks (IPSN)*, 2006, pp. 251–259. [Online]. Available: http://doi.acm.org/10.1145/1127777.1127817

[7] S. C. Geyik and B. K. Szymanski, "Event recognition in sensor networks by means of grammatical inference," in *Proceedings of the 28th IEEE International Conference on Computer Communications*, 2009, pp. 900–908. [Online]. Available: http://www.cs.rpi.edu/~geyiks/geyik_infocom09.pdf

[8] A. S. Ogale, A. Karapurkar, and Y. Aloimonos, "View-invariant modeling and recognition of human actions using grammars," in *Workshop on Dynamical Vision*, 2005, pp. 115–126. [Online]. Available: http://www.cs.umd.edu/~ogale/papers/actionsICCV05WDM.pdf

[9] A. Jagota, R. B. Lyngsø, and C. N. S. Pedersen, "Comparing a hidden markov model and a stochastic context-free grammar," in *Proceedings of the First International Workshop on Algorithms in Bioinformatics (WABI)*, 2001, pp. 69–84. [Online]. Available: http://www.springerlink.com/content/pw22u99rnv8q9p00

[10] Wikipedia, "Catalan number — wikipedia, the free encyclopedia," 2012, [Online; accessed 1-November-2012]. [Online]. Available: http://en.wikipedia.org/w/index.php?title=Catalan_number

[11] J. Earley, “An efficient context-free parsing algorithm,” *Communications of the ACM*, vol. 26, no. 1, pp. 57–61, Jan. 1983. [Online]. Available: http://doi.acm.org/10.1145/357980.358005

[12] A. V. Aho and J. D. Ullman, *The theory of parsing, translation, and compiling.* Prentice-Hall, Inc., 1972.