Method to Identify and Improve Data Quality on IoT Data at WinAd Sites

**Abstract**

In the Internet of things (IoT), data gathered from a global-scale deployment of smart-things are the base for making intelligent decisions and providing services. This paper focuses on the IoT data from a Wind Site with many Wind Turbines and aims to implement a method for detecting and tracking data quality (DQ) of the IoT data. The paper aims to identify the definitions of the DQ dimensions specific to the domain for Wind Sites. After generating a method to detect data quality a dataset from a live wind site is used to determine the effectiveness of the DQ strategy. Data capture settings are adjusted to understand the tradeoffs between DQ dimensions. Data is also categorized and analyzed in batches to understand the DQ of the captured dataset. The results explore any improvements or actionable information that can be used to improve the DQ. In conclusion the possibility of future research and enhancements is presented.

Index

Introduction:

Experiment:

Capture a data set from a scada system for a given set of wind turbine tags and break down the raw measurement data. Use the raw data as a bench mark and establish techniques to measure the data quality dimensions for accuracy.

Discover ways to track the data and then record the data dimensions needed for the system.

Research

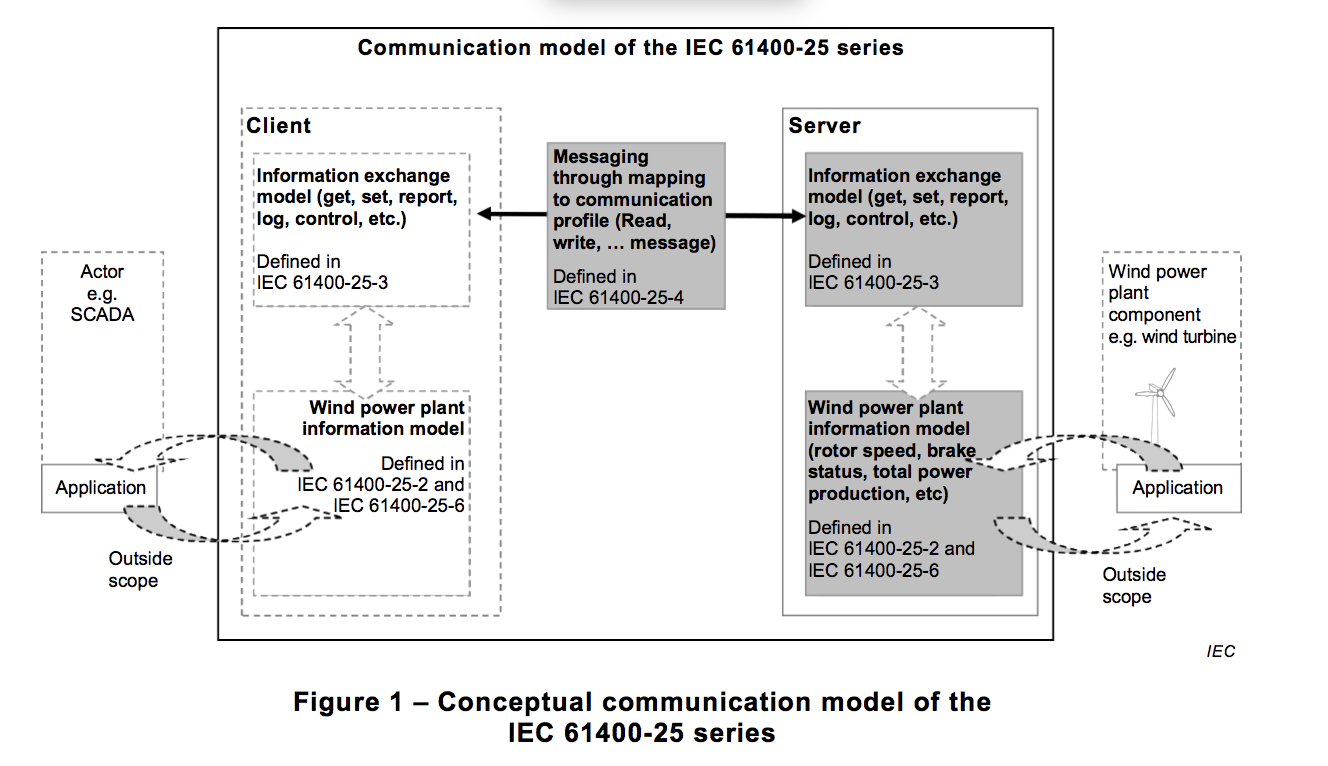
1. SCADA Systems
2. OPC server (opc protocols)
3. How a turbine is set up at a windsite
4. How the data is collected from each instrument?
5. Understanding the tags , identifying Data quality dimensions for each of these tags
   1. IEC standard for wind turbine tags?

[International standard](https://en.wikipedia.org/wiki/International_standard) [**IEC**](https://en.wikipedia.org/wiki/International_Electrotechnical_Commission)**61400-25** (Communications for monitoring and control of wind power plants, TC 88) provides uniform information exchange for monitoring and control of wind power plants. This addresses the issue of proprietary communication systems utilizing a wide variety of protocols, labels, semantics, etc., thus enabling one to exchange information with different wind power plants independently of a vendor. It is a subset of [IEC 61400](https://en.wikipedia.org/wiki/IEC_61400); a set of standards for designing wind turbines.[[1]](https://en.wikipedia.org/wiki/IEC_61400-25#cite_note-1)

The IEC 61400-25 standard is a basis for simplifying the roles that the [wind turbine](https://en.wikipedia.org/wiki/Wind_turbine) and [SCADA](https://en.wikipedia.org/wiki/SCADA) systems have to play. The crucial part of the [wind power plant](https://en.wikipedia.org/wiki/Wind_power_plant) information, information exchange methods, and communication stacks are standardized. They build a basis to which procurement specifications and contracts could easily refer.

The standard has specified five mapping (IEC 61400-25-4) to communication protocol stacks in order to address the real wind power business needs for communication. The mappings specified in the part of IEC 61400-25 comprises a mapping to [SOAP](https://en.wikipedia.org/wiki/SOAP)-based [web services](https://en.wikipedia.org/wiki/Web_service), [OPC](https://en.wikipedia.org/wiki/OLE_for_process_control)/[XML](https://en.wikipedia.org/wiki/XML)-[DA](https://en.wikipedia.org/wiki/OPC_Data_Access), [IEC 61850](https://en.wikipedia.org/wiki/IEC_61850)-8-1 [MMS](https://en.wikipedia.org/wiki/Manufacturing_Message_Specification), [IEC 60870-5](https://en.wikipedia.org/wiki/IEC_60870-5)-104 and a mapping to [DNP3](https://en.wikipedia.org/wiki/DNP3).

* EC 61400-25-1 — Wind turbines - Part 25-1: Communications for monitoring and control of wind power plants - Overall description of principles and models [Published]
* IEC 61400-25-2 — Wind turbines - Part 25-2: Communications for monitoring and control of wind power plants - Information models [Published]
* IEC 61400-25-3 — Wind turbines - Part 25-3: Communications for monitoring and control of wind power plants - Information exchange models [Published]
* IEC 61400-25-4 — Wind turbines - Part 25-4: Communications for monitoring and control of wind power plants - Mapping to communication profile
  + Mapping to [SOAP](https://en.wikipedia.org/wiki/SOAP)-based [Web Service](https://en.wikipedia.org/wiki/Web_Service) [Published]
  + Mapping to [MMS](https://en.wikipedia.org/wiki/Manufacturing_Message_Specification) [Published]
  + Mapping to [OPC](https://en.wikipedia.org/wiki/OLE_for_process_control) [XML](https://en.wikipedia.org/wiki/XML) [Published]
  + Mapping to [IEC 60870-5-104](https://en.wikipedia.org/wiki/IEC_60870-5-104) [Published]
  + Mapping to [DNP3](https://en.wikipedia.org/wiki/DNP3) [Published]
* IEC 61400-25-5 — Wind turbines - Part 25-5: Communications for monitoring and control of wind power plants - Conformance testing [Published]
* IEC 61400-25-6 — Wind Turbines - Part 25-6: Communications for monitoring and control of wind power plants - Logical node classes and data classes for condition monitoring [Published]



1. Some kind of optimization with given constraints
   1. What are my constraints?
      1. Network bandwidth
      2. Get network bandwidth using a tool to measure the amount of traffic on a network card given the architecture of the system? Get the network traffic running on the SCADA machine?

Adelaide :

100m Rotor, Salem pitch controller, 80m tall, MarkVI turbine controller, GE ESS v44.76.00C, WindSCADA 10.0 SP2

Todo: convert these tags into pi tags and other system tags to understand what they are.

Todo: pull in the pi tags from Adelaide that are not on this list.

Can there be a IEC standard naming given to these tags?

Can the behavior of these tags be constrained by technical documentation?

Do the meanings of these tags change across different versions of GE tags?

|  |  |
| --- | --- |
| AI\_CuTorqueAct |  |
| AI\_GenSpdProximitySensor | Sensor data |
| AI\_In\_GridMonAppPowerAct | Power Meter Apparent Power |
| AI\_In\_GridMonReacPowerAct | Power Meter Reactive power |
| AI\_In\_GridMonRealPowerAct | Power Meter Real Power |
| AI\_in\_HubAmbientTemp | Sensor in Hub for temp |
| AI\_In\_MotorCurrentA1 | Current Sensor winding A1 |
| AI\_In\_MotorCurrentA2 | Current Sensor winding A2 |
| AI\_In\_MotorCurrentA3 | Current Sensor winding A3 |
| AI\_In\_TbGbxBearingFastShaftA |  |
| AI\_In\_TbGbxOilTempSump |  |
| AI\_In\_TbNacelleTemperature |  |
| AI\_In\_TbWindSpeed\_NRG | Anemometer Sensor NRG Style |
| AI\_In\_TbWindSpeed\_Thies2D | Anemomentor sensor Thies2d Style |
| AI\_In\_TbWindSpeed\_TVS8002 | Anemometer sensor TVS8002 style |
| AI\_In\_TowerAccelForeAftRaw | Accelerometer foreward and aft |
| AI\_In\_TowerAccelSideSideRaw | Acelleraomter side |
| AI\_In\_WBDelSpdPwr |  |
| AO\_Cu\_TorqueDem |  |
| B\_DynCtl\_WindSpdHigh\_Exit |  |
| B\_In\_WETACommOk |  |
| Curt\_GbxHeatupCurtLossP |  |
| Curt\_GbxHeatupCurtTime |  |
| DI\_House\_WemaFpgaPicChecksumOk |  |
| DI\_In\_BBCtrl\_ValidWindSpeed |  |
| DI\_In\_DtaCommunWEMA\_Ok |  |
| DI\_In\_DtaWEMA28VOk |  |
| DI\_In\_TbCommunWETA\_Ok |  |
| DI\_In\_TbGbxBypassFilterClogged |  |
| DI\_In\_TbGbxBypassFilterPLow |  |
| DI\_In\_TbGbxLubeOilPressureOk |  |
| DI\_In\_TbGbxOilFilterClogCrude |  |
| DI\_In\_TbGbxOilFilterClogFine |  |
| DI\_In\_TbGbxOilLevelLow |  |
| DI\_In\_TbGbxOilParticleCounter |  |
| DI\_In\_TbGbxOilParticleCounterOk |  |
| DI\_In\_TbSwitchGbxOilOverTemp |  |
| DI\_In\_TbWETA28VOk |  |
| DI\_In\_WECA28VOk |  |
| DI\_In\_WETACommOk |  |
| DI\_In\_WETAP24\_OK |  |
| DI\_In\_WETAP24A\_OK |  |
| DI\_In\_WETAP24D\_OK |  |
| DI\_In\_WETAP5\_OK |  |
| DI\_In\_WETAP5A\_OK |  |
| DI\_In\_WETAP5D\_OK |  |
| DistributedIO.WECA\_21.LINK\_OK\_WECA\_R |  |
| DistributedIO.WECA-21.LINK\_OK\_WECA\_R |  |
| DistributedIO.WEMA\_22.LINK\_OK\_WEMA\_R |  |
| DistributedIO.WEMA-22.LINK\_OK\_WEMA\_R |  |
| DistributedIO.WEMA-278.LINK\_OK\_WEMA\_R |  |
| DistributedIO.WEMA-534.LINK\_OK\_WEMA\_R |  |
| DistributedIO.WEMA-790.LINK\_OK\_WEMA\_R |  |
| DistributedIO.WEPA-23.LINK\_OK\_WEPA\_R |  |
| DistributedIO.WEPA-24.LINK\_OK\_WEPA\_R |  |
| DistributedIO.WEPA-25.LINK\_OK\_WEPA\_R |  |
| DistributedIO.WEPA-279.LINK\_OK\_WEPA\_R |  |
| DistributedIO.WEPA-280.LINK\_OK\_WEPA\_R |  |
| DistributedIO.WEPA-281.LINK\_OK\_WEPA\_R |  |
| DistributedIO.WEPA-535.LINK\_OK\_WEPA\_R |  |
| DistributedIO.WEPA-536.LINK\_OK\_WEPA\_R |  |
| DistributedIO.WEPA-537.LINK\_OK\_WEPA\_R |  |
| DistributedIO.WEPA-791.LINK\_OK\_WEPA\_R |  |
| DistributedIO.WEPA-792.LINK\_OK\_WEPA\_R |  |
| DistributedIO.WEPA-793.LINK\_OK\_WEPA\_R |  |
| DistributedIO.WETA\_26.LINK\_OK\_WETA\_R |  |
| DistributedIO.WETA-26.LINK\_OK\_WETA\_R |  |
| DistributedIO.WETA-282.LINK\_OK\_WETA\_R |  |
| DO\_Gbx\_GbxBypassFiltrationOn |  |
| DO\_Gbx\_TbGbxStandbyHeaterOn |  |
| DO\_Gbx\_TbHeatingStrapsOn |  |
| DO\_Gbx\_TbLubePumpMotorHeater |  |
| DO\_Gbx\_TbLubOilCoolFanLow |  |
| DO\_Gbx\_TbLubOilPumpHigh1 |  |
| DO\_Gbx\_TbLubOilPumpHigh2 |  |
| DO\_Gbx\_TbLubOilPumpLow |  |
| DO\_Gbx\_TbOilCoolerMotorHeater |  |
| DO\_Gbx\_TbOilCoolFanHigh |  |
| DO\_House\_WecaLnkOkWemaEgd |  |
| DrvTr\_DrvTrOscAccel |  |
| DynCtl\_ALC\_LoadMeas1 |  |
| DynCtl\_ALC\_LoadMeas2 |  |
| DynCtl\_ALC\_LoadMeas3 |  |
| DynCtl\_ALC\_LoadMeas4 |  |
| DynCtl\_ALC\_NacellePosition |  |
| DynCtl\_Blad1AngleSetpt |  |
| DynCtl\_Blad2AngleSetpt |  |
| DynCtl\_Blad3AngleSetpt |  |
| DynCtl\_BladeAngleSetpt |  |
| DynCtl\_PowerSetpoint |  |
| FaultCode\_A2F |  |
| FaultCode\_A3F |  |
| Gbx\_CoolerFastStageStatus |  |
| Gbx\_CoolerSlowStageStatus |  |
| Gbx\_HeatTapesStatus |  |
| Gbx\_LubPumpFastStageStatus |  |
| Gbx\_LubPumpSlowStageStatus |  |
| Gbx\_Mode |  |
| Gbx\_ParticleCounter |  |
| Gbx\_StandbyHeaterStatus |  |
| In\_GenSpdHiPrecision |  |
| In\_RotorSpd |  |
| In\_TipSpeedRatio |  |
| In\_WIndDirection |  |
| In\_WindSpd |  |
| In\_WindSpd10avg |  |
| MBCEst\_TipSpeedRatio |  |
| OpCrl\_TipSpeedRatio |  |
| OpCtl\_PowerSetpt |  |
| OpCtl\_TurbineOperationState |  |
| OpCtl\_TurbineStatus |  |
| Out\_ExteCurtailSPAverage |  |
| Out\_ExteDetCurtailSPAverage |  |
| Out\_InteCurtailSPAverage |  |
| Out\_InteDetCurtailSPAverage |  |
| Out\_TurbineDownTime |  |
| Out\_TurbineStatusSCADA |  |
| P\_TipSpeedRatioOptimal |  |
| SM\_Curt\_GbxWarmUpInIdling |  |
| SM\_Curt\_HeatUpGbxInLowStage |  |
| SM\_Gbx\_BearingAOverTemp |  |
| SM\_Gbx\_BpFilterPressureLoss |  |
| SM\_Gbx\_ByPassFilterSoiling |  |
| SM\_Gbx\_LuOilLvlLow |  |
| SM\_Gbx\_LuOilPressureLow |  |
| SM\_Gbx\_ManualOperation |  |
| SM\_Gbx\_OilLvlLowAlarm |  |
| SM\_Gbx\_OilOverTemperatureAI |  |
| SM\_Gbx\_OilOverTemperatureDI |  |
| SM\_Gbx\_OilPressureBuildUp |  |
| SM\_Gbx\_OilUnderTemperature |  |
| SM\_Gbx\_PartCnt10minFault |  |
| SM\_Gbx\_PartCnt10minWarn |  |
| SM\_Gbx\_PartCnt1yrFault |  |
| SM\_Gbx\_PartCnt1yrWarn |  |
| SM\_Gbx\_PartCnt7daysFault |  |
| SM\_Gbx\_PartCnt7daysWarn |  |
| SM\_Gbx\_ParticleCounterMalFct |  |
| SM\_Gbx\_SoiledOilFilter |  |
| SM\_Gbx\_SoiledOilFilterWarning |  |
| TD\_TipSpeedRatioGrid |  |
| TD\_TipSpeedRatioTable |  |
| TurbineStatusSCADA |  |
| W\_In\_GbxAuto |  |
| W\_In\_GbxCoolerFastStageOff |  |
| W\_In\_GbxCoolerFastStageOn |  |
| W\_In\_GbxCoolerSlowStageOff |  |
| W\_In\_GbxCoolerSlowStageOn |  |
| W\_In\_GbxHeatTapesHeaterOff |  |
| W\_In\_GbxHeatTapesHeaterOn |  |
| W\_In\_GbxLubPumpFastStageOff |  |
| W\_In\_GbxLubPumpFastStageOn |  |
| W\_In\_GbxLubPumpSlowStageOff |  |
| W\_In\_GbxLubPumpSlowStageOn |  |
| W\_In\_GbxManual |  |
| W\_In\_GbxStandbyHeaterOff |  |
| W\_In\_GbxStandbyHeaterOn |  |
| WIND\_DEV\_10SEC |  |
| WIND\_DEV\_1SEC |  |

**Description:**

1. IoT Data Characteristics
2. Definition of DQ and DQ Dimensions
3. DQ Dimensions for domain-specific application for Smart Grid
4. Finding overlap of DQ Dimensions for Smart Grid and Applying it to Wind Sites
5. Factor affecting DQ in IoT and their impact
   1. Network Bandwidth
   2. Remote Location
   3. Limited Data Storage
   4. Slow Networks
   5. Broken Sensors
   6. Connectivity
   7. Processing Errors
6. DQ Dimensions Trade-offs at Wind Sites
   1. Resolution vs. Completeness
   2. Timeliness vs. Accuracy
7. Present network and logical architecture of a wind site and its collection strategy of IoT data from SCADA and other RTU devices into a Database on the Cloud
8. Design DQ Algorithms for Intrinsic Dimensions
   1. Design algorithms to analyzes DQ in the identified Dimensions given the “Raw” sensor data and the captured data. i.e.:
      1. Resolution : Normalized RMSE between Captured Sensor Data (in the cloud) vs. Raw Sensor Data (from local logs)
      2. Precision : Relative Standard Deviation of Captured Sensor Data vs. Raw Sensor Data
      3. Completeness : Number of tags captured vs. Number of tags on sensor
   2. Using the DQ Analysis algorithms , rewrite them in terms of the hyperparameters and, Identify what the data quality would be if the hyperparameters settings are adjusted in the capturing stage.
      1. If capturing at lower frequency than sensor data
      2. If capturing less data than sensor data
      3. If data is filtered based on deadbands
9. Design DQ Algorithms for Contextual Dimensions
   1. Understand the contextual features of the data
      1. Identify cross-corelated data (i.e. Reactive Power, Active Power, Frequency)
      2. Identify continuous data (temperature, windspeed, etc.)
      3. Identify smooth variation data (humidity, pressure , etc.)
      4. Identify periodic data
   2. Design algorithms to take advantage of the domain specific data context and provide outlier detection to improve data quality analysis
      1. Cross-correlated data should have a correlation
      2. Continuous, smooth variation data should not have jumps
         1. Detecting sharp slopes inside small windows
      3. Periodic data should be able to show periodicity in a given amount of time
         1. Using sampling methods for periodicity of high frequency
         2. Attempt to employ Regression methods to identify periodicity outside of the given dataset (for large periodic data)
10. Given a capture device setting and a batch of data from IoT Data set analyze the data using the DQ Intrinsic and Contextual Dimension algorithms and determine the overall quality of the data as well as granular quality of the sensor data
    1. A data capturing device “Data Logger“, has settings that can effect the DQ
    2. Given the settings determine the effect it has to the sites DQ when capturing the data
    3. With a large batch of data from the site apply the algorithms to the large dataset to determine its data quality. Does the overall data quality change based on the size of the batch? How much data do I need to stabilize the overall data quality measurement.
11. Propose methods to adjust trade-offs
    1. By how much do I need to adjust the settings in the Data Logger to increase the Data Quality?

Appendix

## Data

* Metrics
  + Time series, Time-stamped observations
  + Context
* Static Attributes
  + Units
    - Bytes
    - Requests
  + Data Type
    - Counter
    - Gauge
  + Granularity
  + Raw or Cooked
* Dynamic Attributes
  + Slowly Changing Dimensions (SCD’s)
    - Not Constant, but not always changing
    - Hosts in cluster
    - Number of replicated microservices, etc.
* Collections
  + Entities/Elements
    - Virtual or Physical
    - Dynamic attributes as time series represented as states as it moves through time
    - Example, a javamethodcall over time
  + Relationships
    - Containment, ie. Load Balanced Cluster
    - Sequencing, “Workflow”
  + Interactions
    - Business level work flow
    - Servers, to latency, to customer, to profit

## Types

* Point
  + A point anomaly is an observation that is unusual when compared with all the rest of available observations
* Contextual
  + A contextual anomaly is an observation that is unusual in a certain context but not in other contexts
* Collective
  + A collective anomaly occurs when a collection of related data instances is anomalous with respect to the entire data set.
  + Correlation of multivariate variables

|  |  |
| --- | --- |
| Anomaly type | Detection Data Requirement |
| Point | Uni-variate |
| Contextual | Uni-variate + contextual attributes |
| Collective | Multi-variate+ contextual attributes |

## Techniques

* Deterministic
  + Dashboards
    - Do not scale with growth of metrics
  + Static thresholds
    - Easy to set, but becomes manual
    - Sudden changes above or below thresholds cause false flags
    - Anomalies hidden in cumulative data
  + Transformations (Uni-variate)
    - Delta: raw[n] -> raw [n] – raw[n-1]
    - Rate: raw[n] -> raw [n]/time
    - Scale: raw[n] -> raw [n] \* constant
    - Min: raw[n] -> min(raw)
    - Max: raw[n] -> max(raw)
    - RHMAX: raw[n] -> raw [n]/ max(raw)
    - Summary: raw[n] -> (min,25pct,50pct,mean,75pct,max)
    - Allows for observations in other dimensions
* Statistical
  + Correlation Models
    - Probabilistically stable
    - 1 to -1 correlations between different values
    - As one metric moves the other metric moves the same way(1), or in the opposite (-1)
    - Pearson product moment coefficient of correlation for two metrics
  + Machine learning
    - Learning phase
      * Learning from test data
      * Consume test metrics and provide parameters
    - Detection phase
      * Consume Raw metrics and measure against Parameters
      * Detect anomaly
    - Adjustment
      * Consume raw metrics, measure against parameters
      * Detect anomalies and make corrections to detector

**References :**

1. Aimad Karkouch “Data Quality in Internet of Things: A state of the art servey” 2016

2. [Sabrina Sicaria](https://www.sciencedirect.com/science/article/pii/S0306437916300072" \l "!), [Alessandra R](https://www.sciencedirect.com/science/article/pii/S0306437916300072#!) “A secure and quality-aware prototypical architecture for the Internet of Things” 2016

2. Zita Vale,Hugo Morais,Pedro Faria,Carlos Ramos Distribution system operation supported by contextual energy resource management

based on intelligent SCADA , GECAD – Knowledge Engineering and Decision Support

Research Center, Polytechnic of Porto (IPP), R. Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal

3. K.Metaxiotis A.Kagiannas D.Askounis J.Psarras , Artificial intelligence in short term

electric load forecasting: a state-of-the-art survey for the researcher

Department of Electrical and Computer Engineering, National Technical University

Athens, 9 Iroon Polytechniou str., Zografou 15773, Athens, Greece

Received 12 February 2002, Accepted 10 June 2002, Available online 15 October 2002.