9 November 2016

Third Edition - Dijon - I3M building

Proceeding

Summary

First presentation session - 10h-12h30

T. Hassan - "Hierarchical Multi-Label Classification Using Web Reasoning for Large Datasets"
M. Khokhlova - "3D Visual-based Human Motion Descriptors : A Review"
A. Boscaro - "Automatic processing scheme for low laser invasiveness electro optical frequency mapping mode"
H. Leake Kidane - "NoC based virtualized FPGA as cloud Services"
N. Piasco - "Collaborative localization and formation flying using distributed stereo-vision"

Poster session - 13h30-14h30

R. Marroquin - "WiseNET : smart camera network interacting with a semantic model"
C. Bensekka - "Topological analysis of movement"
B. Li - "Multi-user Interface for Co-located Real-time Collaborative Work with Digital Mock-up"

Second presentation session - 14h30-17h

A. Zawawi Jamaluddin - "A New Method for Omni-RGB+D Camera Rig Calibration and Fusion using Unified Camera Model"
R. Macwan - "Remote Photoplethysmography with Constrained ICA using Autocorrelation as a periodicity measure"
S. Bobbia - "Remote Photoplethysmography Based on Implicit Living Skin Tissue Segmentation"
D. Strubel - "Comparison between genetic algorithm and particle swarm optimization for positionning a set of cameras for video surveillance"
A. Zanzouri Kechiche - "Shape from polarization in the far IR applied to 3D digitization of transparent objects"

Hierarchical Multi-Label Classification Using Web Reasoning for Large Datasets

Thomas Hassan
Le2i UMR6306, CNRS, Arts et Métiers,
Univ. Bourgogne Franche-Comté
thomas.hassan@u-bourgogne.fr

Christophe Cruz
Le2i UMR6306, CNRS, Arts et Métiers,
Univ. Bourgogne Franche-Comté
christophe.cruz@u-bourgogne.fr

Abstract

Determining valuable data among large volumes of data is one of the main challenges in Big Data. We aim to extract knowledge from these sources using a Hierarchical Multi-Label Classification process called Semantic HMC. This process automatically learns a label hierarchy and classifies items from very large data sources. Five steps compose the Semantic HMC process. This paper focuses on the last two steps where new items are classified according to the label hierarchy. The process is implemented in a scalable and distributed platform to process Big Data. The process is evaluated and compared with multi-label classification algorithms from the state of the art dedicated to the same goal where the Semantic HMC approach outperforms state of the art approaches in some areas.

1. Introduction

The item analysis process requires proper techniques for analysis and representation. In the context of Big Data, this task is even more challenging due to Big Data's characteristics. An increasing number of V's has been used to characterize Big Data [1]: Volume, Velocity, Variety and Value. Volume concerns the large amount of data that is generated and stored through the years by social media, sensor data, etc.[1]. Velocity concerns both the production and the process to meet a demand because Big Data is not only a huge volume of data but it must be processed quickly as new data is generated over time. Variety relates to the various types of data composing the Big Data. These types include semi-structured and unstructured data representing 90% of his content such as audio, video and text. Value measures how valuable the information to a Big Data consumer is. Value is the most important feature of Big Data, because the user expects to make profit out of valuable data. As Big Data analysis can be deemed as the analysis of a special kind of data, many traditional data analysis methods used in Data Mining (algorithms for classification, clustering, regression, among others) may still be utilized for Big Data Analysis [1]. Werner et al. [2] propose a method to semantically enrich an ontology used to describe the domain and classify the news articles. This ontology aims to reduce the gap between the expert's perspective and the classification rules representation. To enrich the ontology and classify the documents they uses an out-of-the-box Description Logics (DL) Web Reasoner. Most of these reasoners are sound and complete to high expressiveness, as OWL2 SROIQ (D) expressiveness, but on the other hand they do not scale: these reasoners cannot handle a large amount of data. Our goal is to extend the work in [2] and to exploit value by analyzing Big Data using a Semantic Hierarchical Multi-Label Classification process (Semantic HMC)[3]. The Semantic HMC is based on an unsupervised ontology learning process using scalable Machine-Learning techniques and Rule-based reasoning. The ontology-described knowledge base (Abox+Tbox) used to represent the knowledge in the classification system is automatically learned from huge volumes of data through highly scalable Machine Learning techniques and Big Data Technologies. Semantic HMC proposes five individually scalable steps to reach the aims of Big Data analytics [4]:

- **Indexation** extracts terms from data items and creates an index of data items.
- **Vectorization** calculates the term-frequency vectors of the indexed items.
- **Hierarchization** creates a label taxonomy (i.e. subsumption hierarchy) from term-frequency vectors.
- **Resolution** creates the reasoning rules to relate data items with the labels based on term-frequency vectors.
- **Realization** first populates the ontology with items and then for each item determines the most specific label and all its subsuming labels.

[3] focuses on the two last steps of the Semantic HMC process. It proposes a new process to hierarchically multi-classify items from huge sets of unstructured texts using DL

ontologies and Rule-based reasoning. This paper is an extention of the work presented in [3] and provides extended experiments with quality evaluation and comparison with some multi-label classification algorithms from the state of the art. The rest of the paper covers five sections. The second section presents background and related work. The third section describes the classification process. The forth section describes the process implementation in a scalable and distributed platform to process Big Data. The fifth section discusses the results. Finally, the last section draws conclusions and suggests further research.

2. Related work

2.1. Ontologies in Classification context

Ontologies are recurrently used in classification systems to describe the classification knowledge (labels, items, classification rules) and to improve the classification process. Galinina et al. [5] used two ontologies to represent a classification system: (1) a Domain ontology that is independent of any classification method and (2) a Method ontology devoted to decision tree classification. Beyond domain description, ontologies can be used to improve the classification process. Elberrichi et al. [6] present a two-steps method for improving classification of medical documents using domain ontologies (MeSH - Medical Subject Headings). Their results prove that document classification in a particular area supported by ontology of its domain increases the classification accuracy.

2.2. Web reasoning in Classification context

Reasoning is used at ontology development or maintenance time as well as at the time ontologies are used for solving application problems [7]. In Classification context, Web reasoning can be used to improve the classification process. In [8] authors presents a document classification method that uses ontology reasoning and similarity measures to classify the documents. In [9] authors introduce a generic, automatic classification method that uses Semantic Web technologies to define the classification requirements, perform the classification and represent the results. The proposed generic classifier is based on an ontology, which gives a description of the entities that need to be discovered. In [2] authors uses out-of-the-box reasoning to classify economical documents but their scalability is limited and cannot be used in large datasets as required in Big Data context.

2.3. Discussion

Most work in the literature focus on describing or improving the classification processes using ontologies but do not take advantage of the reasoning capabilities of web reasoning to automatically multi-classify the items. However as Semantic Web is growing, new high-performance Web Scale Reasoning methods have been proposed [10]. Rule-based reasoning approach allows the parallelization and distribution of work by large clusters of inexpensive machines by programming models for processing and generating large data sets such as Map-reduce[11]. Web Scale Reasoners [10] however, instead of using traditional DL approaches, use entailment rules for reasoning over ontologies. Web-Scale Reasoners based on Map-reduce programming model like WebPie [10] outperforms all other published approaches in an inference test over 100 billion triples [10]. In [12] authors describe a kind of semantic web rule execution mechanism using MapReduce which can be used with OWL-Horst and with SWRL rules. To the extent of our knowledge, a classification process to automatically classify text documents in Big Data context by taking advantage of ontologies and rule-based reasoning to perform the classification is novel.

3. Hierarchichal Multi-Label Classification

In [13], the authors describe in detail the first three steps (Indexation, Vectorization and Hierarchization) of the classification process. Beyond learning the label hierarchy, the process aims to learn a classification model based on a DL ontology presented in [3]. The following subsections describe the last two steps of the process, i.e. how the rules used to classify the items are created and how items are classified using Rule-based Web Reasoning.

3.1. Resolution

The resolution step creates the ontology rules used to relate the labels and the data items, i.e. it establishes the conditions for an $item_i$ to be classified as $label_j$. The rules will define the necessary and sufficient terms of an item $item_i$ be classified as $label_j$. The rules creation process uses thresholds as proposed in [2] to select the necessary and sufficient terms. The main difference with this method is that instead of translating the rules into logical constraints of an ontology captured in Description Logic, these rules are translated in the Semantic Web Rule Language (SWRL). The main interest in using SWRL rules is to reduce the reasoning effort, thus improving the scalability and performance of the system. In the Vectorization step[13], a term co-occurrence frequency matrix $cfm(term_i, term_j)$ is created to represent the co-occurrence of any pair of terms in the collection of items C. Let $P(term_j|term_i)$ be the conditional proportion (number) of the items from collection C common to $term_i$ and $term_j$, in respect to the number of items in $term_j$ such that:

$$P_C(term_i|term_j) = \frac{cfm(term_i, term_j)}{cfm(term_j, term_j)} \qquad (1)$$

Two thresholds are defined:

- Alpha threshold (α) such that $\alpha < P_C(term_i|term_j)$, where $term_i \in Label$ and $term_j \in Term$.
- Beta threshold (β) such that $\beta \leq P_C(term_i|term_j) \leq \alpha$, where $term_i \in Label$ and $term_j \in Term$.

These two thresholds are user-defined with a range of $[0, 1]$. Based on these thresholds, two sets of terms are identified (Fig.1):

- Alpha set ($\omega_\alpha^{(term_i)}$) is the set of terms for each label such that:

$$\omega_\alpha^{(term_i)} = \{term_j | \forall term_j \in Term : P_C(term_i|term_j) > \alpha\} \quad (2)$$

  i.e. is the set of terms $term_j$ that co-occur with $term_i \in Label$ with a co-occurrence proportion higher than the threshold α.

- Beta set ($\omega_\beta^{(term_i)}$) is the set of terms for each label such that:

$$\omega_\beta^{(term_i)} = \{term_j | \forall term_j \in Term : \beta \leq P_C(term_i|term_j) \leq \alpha\} \quad (3)$$

  i.e. is the set of terms that co-occur with $term_i \in Label$ with a co-occurrence proportion higher or equal than the threshold β and lower than the threshold α.

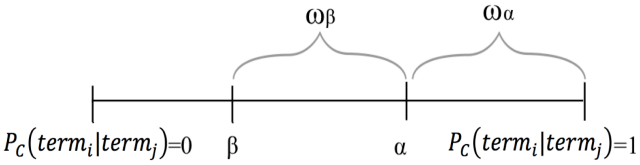


Figure 1. Alpha and Beta sets

.

If the item has at least one term in $\omega_\alpha^{(term_i)}$ it is classified with $term_i$, $term_i \in Label$. For each term that complies with the above rule, a SWRL rule is created.

If the item has at least δ terms in $\omega_\beta^{(term_i)}$, it is classified with $term_i$, $term_i \in Label$. One SWRL rule is generated for each combination of $term_j \in \omega_\beta^{(term_i)}$ where the number of combined terms is at least $\delta = \lceil |\omega_\beta^{(term_i)}| * p \rceil$, and $0 \leq p \leq 0.5$. The set of generated beta rules is the combination C_n^m of m terms of a larger set of n elements. Regarding our approach, n is the number of possible terms $|\omega_\beta^{(term_i)}|$, and m the minimum number of terms δ in each rule (e.g. $C_{20}^{10} = 184756$). In order to limit the number of rules for each label we fix the value of $n \leq 10$. The terms are selected by ranking the terms in $\omega_\beta^{(term_i)}$ using the conditional proportion $P_C(term_i|term_j)$ as the ranking score. Example Alpha and Beta SWRL rules are depicted in Table 1.

Notice that the rules that encompass more than δ terms are not necessary because the combination of any δ terms is sufficient to classify the item.

In non-empty alpha and beta category, beta and alpha rules are both considered. Alpha rules are evaluated as presented in the empty beta category. Beta rules are evaluated as presented in the empty alpha category but with a value $q = p * 2$ because beta rules are, by definition, less relevant than alpha rules. It corresponds to $\delta = \lceil |\omega_\beta^{(term_i)}| * q \rceil$, with $0 \leq q \leq 1$ and $q = p * 2$.

3.2. Realization

The realization step includes two sub-steps: population and classification. The ontology-described knowledge base is populated with new items and their relevant terms at the assertion level (Abox). Each item is described with a set of relevant terms $\omega_\gamma^{(item_i)}$ such that:

$$\omega_\gamma^{(item_i)} = \{term_j | \forall term_j \in Term \wedge \gamma < tfidf_{(item_i, term_j, C)}\} \quad (4)$$

where γ is the relevance threshold, $\gamma < tfidf_{(item_i, term_j, C)}$, $term_j \in Term$, $item_i \in Item$ and $tfidf$ as calculated in the Vectorization step.

The classification sub-step performs the multi-label hierarchical classification of the items. Rule-based reasoning applies exhaustively a set of rules to a set of triples (i.e. the data items) to infer conclusions [14], i.e. the item's classifications.

The rule-based inference engine uses rules to infer the subsumption hierarchy (i.e. concept expression subsumption) of the ontology and the most specific concepts for each data item. This leads to a multi-label classification of the items based in a hierarchical structure of the labels (Hierarchical Multi-label Classification).

4. Implementation

The process is implemented as a combination of available Java libraries that natively support parts of the process. In the first three steps (indexation, vectorization and hierarchization) of the Semantic HMC process, Big Data technologies are used, including MapReduce [11]. The MapReduce algorithms are deployed on a Hadoop cluster https://hadoop.apache.org/. The next subsections describe the implementation details of each step of the classification process.

4.1. Resolution

The resolution process creates the ontology rules used to relate the labels and the data items. The rule creation process is divided in a sub-process for each $label_i \in Label$. In

Table 1. Generated Rules Examples

Alpha rules
$Item(?it), Term(?t_1), Label(?t_1), hasTerm(?it, ?t_1) \rightarrow isClassified(?it, ?t_1)$
Beta rule
$Item(?it), Term(?t_1), Term(?t_2), Label(?t_3), hasTerm(?it, ?t_1), hasTerm(?it, ?t_2) \rightarrow isClassified(?it, ?t_3)$

each sub-process $\omega_{\alpha}^{(label_i)}$ and $\omega_{\beta}^{(label_i)}$ sets are calculated using the co-occurrence matrix, then classification rules are created for each label. Exploiting a huge co-occurrence matrix to create the ontology rules is a very intensive task, thus this process is also distributed to several machines in the MapReduce paradigm. A MapReduce job creates the rules from the co-occurrence matrix. Following rule generation in MapReduce, the rules are serialized in SWRL language and stored in the ontology-described knowledge base using the OWL-API library. The generated rules along with the label hierarchy are used in the Realization process to classify new items.

4.2. Realization

The realization step populates the ontology and performs the multi-label hierarchical classification of the items. First the ontology is populated with new items and the most relevant terms to describe each document in an assertion level (Abox). To store, manage and query the ontology-described knowledge base (Tbox+Abox) a triple-store is used. Because highly expressive forward chaining description logics reasonners do not scale well and, in our preliminary prototype we decided to adopt the classification at query time approach by using a triple-store with a backward-chaining inference engine. The OWL-API library is used to populate the OWL ontology with new items. A scalable triple-store called Stardog (http://docs.stardog.com) is used to store and query the ontology-described Knowledge Base (Tbox+Abox). Stardog is also used to perform reasoning by backward-chaining inference as well as SWRL rules inference. The rule selector was developed in java, and interacts with Stardog to optimize the query performance.

5. Experiments

In this section, a quality evaluation of the Semantic HMC is depicted. This evaluation focuses on classification accuracy, and complements the performance evaluation depicted in [3]. Finally we discuss the obtained results regarding some algorithms from the state of the art in Hierarchical Multi-Label Classification.

5.1. Quality Evaluation

In this subsection we evaluate the classification performance of the Semantic HMC process for unstructured text classification in a Big Data context. First the dataset, the test environment and the experimental settings used to evaluate the process are described. Then the experimental results are presented and discussed. The evaluation is done using a pre-labeled dataset, composed of training and test data. The training set is used to learn hierarchical relations between the pre-defined labels and classification rules. The test set is used to calculate the classification performance of the algorithm based on standard quality measures. To be able to compare our approach with state-of-the-art, we use a pre-defined set of labels instead of automatically learned labels as it is described in [3].

5.1.1 Delicious dataset

The Delicious dataset[1] is used to perform this evaluation. This dataset is composed of labeled textual data from web pages extracted from the Delicious social bookmarking website[15]. Table 2 shows the dataset specifications. The Delicious dataset was chosed because contains very few features (words) compared to the number of labels, rendering accurate classification difficult [16]. Also, it has been used to evaluate several multi-label classification systems, thus it provides a good baseline to compare our approach.

Table 2. Delicious dataset specifications

$\|Train\|$	$\|Test\|$	$\|Labels\|$	$\|Terms\|$
12,910	3,181	983	500

5.1.2 Measures

Specific evaluation metrics to multi-label learning are proposed in literature and generally categorized into two groups : example-based metrics and label-based metrics[17]. We use a label-based metric to evaluate the Semantic HMC. In label-based metrics the micro-averaged as well as the macro-averaged precision and recall are used : the learning system's performance is evaluated on each class label separately, and then the macro/micro-averaged value across all class labels is returned. These measures are calculated as in [18].

5.1.3 Results

The Hierarchization phase of the Semantic HMC process automatically generates a hierarchical relations between labels. This hierarchy, along with the classification rules cre-

[1] http://mulan.sourceforge.net/datasets-mlc.html

ated in the Resolution step are used to perform hierarchical multi-label classification. Figure 2 shows a sample of the hierarchichal relations ($skos : hasBroaderRelation$) between labels automatically created for the Delicious dataset. The set of parameters used to create the hierarchy and classification rules is described in table 3. This parameters can have a high impact in the quality of the results. The Top and Bottom Tresholds are used to calculate the hierarchichal relations between labels as defined in [13].

Table 3. Execution Settings for Delicious Dataset

Parameter	Step	Value
Top Threshold	Hierarchization	50
Bottom Threshold	Hierarchization	40
Alpha Threshold	Resolution	20
Beta Threshold	Resolution	10
Term ranking (n)	Resolution	5
p	Resolution	0.25
Term Threshold (γ)	Realization	2

Table 4 shows the results obtained by the Semantic HMC process on the Delicious dataset.

Table 4. Quality results for the Delicious Dataset

	Precision	Recall	F1-measure
Micro	0.284	0.74	0.410
Macro	0.0676	0.178	0.0979

5.2. Comparison with the state of the art

Table 5 shows the Macro-F1 measure and Micro-F1 measure obtained on the Delicious dataset. The results of the proposed process (SHMC) are compared with several state-of-the-art approaches results with the same dataset[18][19][16]. In Table 5, it is observed that the

Table 5. Performance of various algorithms on the Delicious dataset

Algorithm	Macro F1	Micro F1
SHMC	0.0979	**0.410**
CGS_p	0.10378	0.29740
TNBCC	0.0880	N/A
Path-BCC	0.084	N/A
BR	0.096	0.234
CC	0.100	0.236
HOMER	0.103	0.339
ML-kNN	0.051	0.175
RFML-C4.5	**0.142**	0.269
RF-PCT	0.083	0.248

Semantic HMC approach outperforms state-of-the-art approaches in micro F1-measure, while the macro F1-measure is comparable to most other approaches. These results show that the classification performance of our ontology-based approach is comparable to the performance of the selected algorithms from the state-of-the-art in machine learning.

6. Conclusions

This paper describes an unsupervised hierarchical multi-label classification process from unstructured text in the scope of Big Data. Following the performance evaluation depicted in [3], a quality evaluation is depicted, comparing the classification accuracy of the Semantic HMC with several approaches from the state-of-the-art. The experiment shows that the classification performance of the Semantic HMC process that uses ontologies and rule-based reasoning to classify unstructured text documents is comparable to the performance of algorithms from the state-of-the-art in machine learning field. Also, unlike most approaches from the data-mining field, the ontology-based approach provides human-readable explanations of the classifications, that can be used to monitor the classification process by experts. Our current work is twofold: (1) the application of the process to domain-specific data and (2) the maintenance of the classification model regarding a stream of data in a Big Data Context.

References

[1] "Big data: A survey", *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, Jan. 2014, ISSN: 1383469X. DOI: 10.1007/s11036-013-0489-0 (cit. on p. 1).

[2] D. Werner, N. Silva, C. Cruz, and A. Bertaux, "Using DL-reasoner for hierarchical multilabel classification applied to economical e-news", in *Proceedings of 2014 Science and Information Conference, SAI 2014*, 2014, pp. 313–320, ISBN: 9780989319317. DOI: 10.1109/SAI.2014.6918205 (cit. on pp. 1, 2).

[3] R. Peixoto, T. Hassan, C. Cruz, A. Bertaux, and N. Silva, "An unsupervised classification process for large datasets using web reasoning", in *SBD'16: Semantic Big Data Proceedings*, ACM, Ed., San Francisco (CA), USA, 2016 (cit. on pp. 1, 2, 4, 5).

[4] T. Hassan, R. Peixoto, C. Cruz, A. Bertaux, and N. Silva, "Semantic HMC for big data analysis", in *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, 2015, pp. 26–28, ISBN: 9781479956654. DOI: 10.1109/BigData.2014.7004482. arXiv: 1412.0854 (cit. on p. 1).

[5] A. Galinina and A. Borisov, "Knowledge modelling for ontology-based multiattribute classification system", *Applied Information and Communication ...*, pp. 103–109, 2013 (cit. on p. 2).

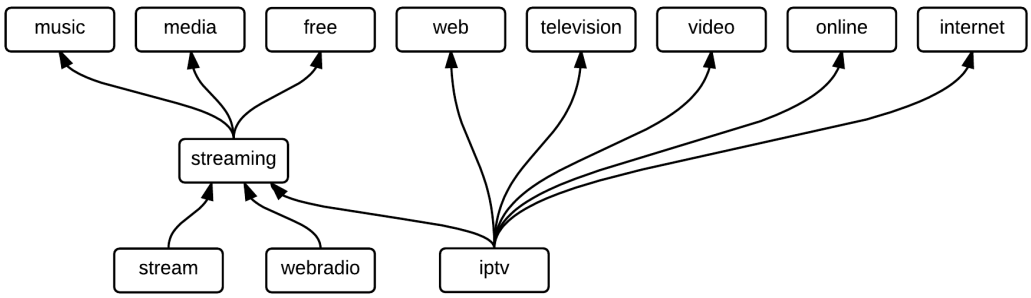


Figure 2. Automatically generated hierarchy from Delicious dataset (sample)

[6] Z. Elberrichi, B. Amel, and T. Malika, "Medical Documents Classification Based on the Domain Ontology MeSH", *ArXiv preprint arXiv:1207.0446*, 2012 (cit. on p. 2).

[7] R. Moller and V. Haarslev, *Tableau-Based Reasoning*, 2009. DOI: 10.1007/978-3-540-92673-3{_}23 (cit. on p. 2).

[8] J. Fang, L. Guo, and Y. Niu, "Documents classification by using ontology reasoning and similarity measure", in *Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010*, vol. 4, 2010, pp. 1535–1539, ISBN: 9781424459346. DOI: 10.1109/FSKD.2010.5569338 (cit. on p. 2).

[9] D. Ben-David, T. Domany, and A. Tarem, "Enterprise data classification using semantic web technologies", in *The Semantic Web-ISWC 2010*, ser. ISWC'10, Berlin, Heidelberg: Springer-Verlag, 2010, pp. 66–81, ISBN: 3-642-17748-4, 978-3-642-17748-4 (cit. on p. 2).

[10] J. Urbani, "Three Laws Learned from Web-scale Reasoning", in *2013 AAAI Fall Symposium Series*, 2013, pp. 76–79 (cit. on p. 2).

[11] J. Dean and S. Ghemawat, "MapReduce : Simplified Data Processing on Large Clusters", *Communications of the ACM*, SIGMOD '07, vol. 51, no. 1, L. P. Daniel, Ed., pp. 1–13, 2008, ISSN: 00010782. DOI: 10.1145/1327452.1327492. arXiv: 10.1.1.163.5292 (cit. on pp. 2, 3).

[12] H. Wu, J. Liu, D. Ye, H. Zhong, and J. Wei, "A distributed rule execution mechanism based on MapReduce in sematic web reasoning", *Proceedings of the 5th Asia-Pacific Symposium on Internetware - Internetware '13*, pp. 1–7, 2013. DOI: 10.1145/2532443.2532457 (cit. on p. 2).

[13] R. Peixoto, T. Hassan, C. Cruz, A. Bertaux, and N. Silva, "Semantic HMC: A Predictive Model using Multi-Label Classification For Big Data", in *The 9th IEEE International Conference on Big Data Science and Engineering (IEEE BigDataSE-15)*, 2015, ISBN: 978-1-4673-7952-6. DOI: 10.1109/Trustcom.2015.578 (cit. on pp. 2, 5).

[14] J. Urbani, F. Van Harmelen, S. Schlobach, and H. Bal, "QueryPIE: Backward reasoning for OWL horst over very large knowledge bases", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, ser. ISWC'11, vol. 7031 LNCS, Berlin, Heidelberg: Springer-Verlag, 2011, pp. 730–745, ISBN: 9783642250729. DOI: 10.1007/978-3-642-25073-6{_}46 (cit. on p. 3).

[15] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective and efficient multilabel classification in domains with large number of labels", in *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*, 2008, pp. 30–44 (cit. on p. 4).

[16] Y. Papanikolaou, T. N. Rubin, and G. Tsoumakas, "Improving gibbs sampling predictions on unseen data for latent dirichlet allocation", *ArXiv preprint arXiv:1505.02065*, 2015 (cit. on pp. 4, 5).

[17] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms", *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014, ISSN: 10414347. DOI: 10.1109/TKDE.2013.39 (cit. on p. 4).

[18] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning", *Pattern Recognition*, vol. 45, no. 9, pp. 3084–3104, 2012 (cit. on pp. 4, 5).

[19] L. E. Sucar, C. Bielza, E. F. Morales, P. Hernandez-Leal, J. H. Zaragoza, and P. Larranaga, "Multi-label classification with bayesian network-based chain classifiers", *Pattern Recognition Letters*, vol. 41, pp. 14–22, 2014, Supervised and Unsupervised Classification Techniques and their Applications, ISSN: 0167-8655 (cit. on p. 5).

3D Visual-based Human Motion Descriptors: A Review

Margarita Khokhlova
Le2i UMR6306, CNRS, Arts et Métiers, Univ. Bourgogne Franche-Comté, France
margarita.khohklova@u-bourgogne.fr

***Abstract*—This paper aims to provide a comprehensive reference source on depth-based human motion descriptors. Motion description is a challenging problem which became popular with recent advances in 3D computer vision. Our goal is to introduce the main trends in human 3D motion descriptor design and evaluation next to presenting a review of recent methods belonging to three application categories: action recognition, gesture recognition and gait assessment.**

***Keywords*-human motion description; motion 3D; action recognition; gesture recognition; gait assessment.**

I. Introduction

Visual or image descriptors are short feature vectors of the salient regions of images and video sequences. Descriptors aim to describe the visual characteristics of present objects such as shape, color, texture or motion and may replace an image as the input to a classifier.

Specially designed and tailored descriptors can also represent motion, which is an essential part of algorithms in rather diverse applications, such as the human activity recognition, gait recognition and analysis, motion tracking, 3D scene reconstruction to name a few examples.

With the advent of low-cost 3D sensing cameras and continued efforts in advanced point cloud processing, 3D perception has gained more importance in the vision domain. 3D sensing devices not only provide the user with a general projection of the 3D world to a 2D image plane as regular cameras, but also acquire the 3D geometry, or depth. The depth images deliver natural surfaces which can be exploited to capture geometrical features of the observed scene with a rich descriptor. Compared with conventional color videos, the additional depth information in RGBD data helps to adjust for different lighting conditions, remove background noise and simplify intra-class motion variations. Therefore, in general, RGBD-based descriptors outperform the RGB-based ones [1].

The standard output of a 3D sensor is a point cloud, which contains points on the scene next to their XYZ coordinates and an optional color tuple. A descriptor specifically designed for a point cloud video sequence could serve as an important basis for motion characterization.

The information extracted from 3D point clouds is predominantly comprised of the shape, color (or intensity) and the spatial relation between cloud points. Shape descriptors are the most popular 3D descriptors for point clouds, and amongst them normal-based descriptors are the most widely used.

Widely applied three dimensional shape descriptors include the Normal Aligned Radial Features (NARF) [2] and the Spin image descriptor [3] whereas for two dimensional planes HOG3D [4], SURF and SIFT [5] are common. The latter may be used to represent a motion trajectory when applied to individual consecutive frames [6]. Many have been implemented by the Point Cloud Library [7].

Finally, 3D motion, sometimes referred to as 4D, combines 3D spatial scenes though time. This data may be described via the calculation of the so-called Motion or Scene Flow, which are build of from 3D velocity vectors. Dense motion flow is information rich, but is computationally intensive to determine and carries a significant memory and storage burden. Thus in recent years many alternative algorithms have been proposed that aim at reducing this burden. This paper aims to review and categorize this existing family of methods, specifically for applications in human motion analysis.

Motion analysis using depth data was extensively reviewed in [8]. However, in this work researchers were focused primarily on activity recognition methods applied to RGBD video sequences omitting (3D) motion descriptors. Therefore, we assumed a need for a specialized review dedicated to 3D descriptors applied to human motion, divided into the following subcategories: Motion Recognition, Gesture Recognition, and Movement Analysis and specifically, Gait recognition.

The remainder of this article is organized as follows. Section 2 presents the requirements and evaluation criteria for motion descriptors. In the second part of Section 2 we introduce the most popular 3D datasets for motion descriptor evaluation. Further in Section 3, 4 and 5 we talk about the common trends and representative methods to describe the motion for 3 applications: action recognition, gesture recognition and gait analysis. We summarize the main approaches and point out their interesting and novel parts. Finally, Section 6 concludes the paper, where we highlight the main existing trends in motion description for point clouds sequences and propose our outlook for the future.

II. Motion descriptor evaluation

A. Motion descriptor characteristics

Common requirements for 3D descriptors are invariance to transformations of the target object, user-independence, robustness to noise and clutter and storage efficiency (i.e., compactness). We propose to compare existing descriptors by a combination of the following characteristics, which highlight descriptor specifics and give an idea in which scenarios their performance is optimal:

Application: Corresponds to the particular purpose for which the descriptor is designed.

Locality: Descriptor can be local (regional) or global based on the features it captures. If the descriptor is local, it is applied to selected points of interest, or in a local support region surrounding a basis point of an object. In this case, before applying the descriptor, potentially interesting points should be identified, preferably, automatically. Rarely researchers try to select the most representative points by applying a descriptor to all the data points of the model and comparing the values to find 'rare' ones [9]. Commonly algorithms may be split in a detector and a descriptor part, where the detector locates regions that are perceived to be interesting or stable and the descriptor encodes a region of interest around the location of the feature. Using motion descriptors for local features generally provides one with invariance to geometric transformations and gives a reduction of the perturbations caused by variations in scale, rotation, and viewpoint. Using local descriptors also allows to reduce the computational complexity of the algorithms. However, local descriptors ignore the global spatial structure information of the scene.

Global descriptor is applied to all points or to points selected by using a linear sampling scheme and is capable of capturing global spatial structure and affiliation.

Analogous to the descriptors for 2D images, descriptors applied locally are more common due to the fact that the cost of globally computing descriptors is usually high.

Dimensionality: Usually directly connected to the size and representativeness of a descriptor. Dimensionality shows how many values are stored in each descriptor array. Dimensionality of a motion descriptor is rarely a significant issue unless the descriptors for very long video sequences should be stored or the rich motion flow based descriptor is used. In this case the general storage burden is predominantly determined by the number of descriptors and not their individual size.

View-invariance and scale-invariance: View-invariance of a descriptor entails its robustness against changes in view-point. Many approaches for 3D shape retrieval and identification transform an object of interest into a canonical pose. Translating the center of gravity of the object into the origin and normalizing the area/volume or radius of the bounding circle/sphere/parallelepiped etc. This operation guarantees view-invariance to a reasonable extend.

Scale-invariance is the invariance of the descriptor for the objects size or the distance of the camera to the object. In Local Spatio-Temporal (LST) descriptors it can be, for example, the fixed support region of the feature point which is not adapted to the linear perspective view variations.

Accuracy: Characteristic which shows how well the proposed descriptor performed for a given task and if it corresponds to the task at hand. In this work, we state the accuracy as reported by the researchers and also the theoretical evaluation of the method proposed. When available, we report the results obtained on a benchmark dataset. However, we should mention that even when an identical dataset is used, the validation method used by each work might differ from the others [10], so a direct comparison is not always possible.

Computational complexity : The calculation resources required by an algorithm. Sometimes computational complexity is reported by the authors or could be approximated by the specifics of the algorithm.

Classifier: The algorithm that implements the recognition task in the application. The use of a different classifier can actually change the final recognition score significantly, however, with a set of discriminative features even a very simple classification method like linear regression can show good results. It is very common to use the Bag-of-Words model for classification tasks ([11], [12], [13], [14], [6]) and in this case the model defines the used classifier.

Dataset: Motion descriptors reviewed in this work were proposed for a particular application and evaluated using task corresponding test data. The choice of a dataset obviously affects the reported accuracy of an algorithm. Datasets will be evaluated on their difficulty and their data scope.

B. Datasets for the motion descriptors

When proposing a new algorithm, its performance usually needs to be evaluated in comparison to existing ones for a given application. Construction and annotation of a new database is often a long and arduous process, therefore it is preferable to do only do this when an existing benchmark can not be used.

We propose a list of popular Benchmark datasets, most widely used for 3D human motion descriptor evaluation next to their specifics and designation for a particular task.

MSR Action3D: This is the most used RGBD human action-detection and recognition dataset [11], [15], [16], [17], [18]. Mainly because this is one of the first RGBD datasets capturing motions (dated 2010) and it contains the biggest amount of different actions. The MSR Action3D Dataset [12] consist of 20 action types performed by 10 subjects 2 or 3 times. The actions are: high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw an x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up & throw. The resolution of the video is not very high, namely 320x240. The data was recorded with a depth sensor similar to the Kinect device. It is a challenging dataset to test an algorithm on and to compare the results with earlier proposed methods.

MSRGesture3D: This dataset was captured by a Kinect device. There are 12 dynamic American Sign Language gestures performed 2-3 times by 10 people. The hand sign language represents a concentrated entity in motion, combining both the overall movement of hands and the inner variability of fingers. The hand portion (above the wrist) is segmented. An alternative to this dataset with a

less number of gestures is the Sheffield Kinect Gesture (SKIG) dataset [19], which captures forearm gestures under different hand poses, background and illumination variations from 6 subjects. This makes SKIG more applicable to a real-life scenario than the MSR Gesture 3D dataset. However, the latter remains the most widely used for gesture recognition research [20], [21], [22].

MSR Daily Activity 3D: Also captured using a Kinect device, it includes 16 activities performed by 10 subjects 2 times, in a standing and a sitting position. The actions are: drink, eat, read a book, call a cellphone, write on a paper, use a laptop, use a vacuum cleaner, cheer up, sit still, toss a paper, play a game, lie down on a sofa, walk, play a guitar, stand up, sit down. There is a sofa in the scene. RGB and depth channels are recorded, and also the skeleton joint positions are extracted. However, the RGB channel and depth channel are recorded independently, so they are not strictly synchronized. The dataset is more challenging than MSR Action3D, because it represents natural everyday activities, which are harder to distinguish. MSR Daily Activity 3D is a good choice to evaluate a real-life scenario dedicated application and compare the results with other algorithms [14], [16], [15].

Berkeley MHAD: The Berkeley Multimodal Human Action Database (MHAD) [23] is a complete and general purpose dataset which consist of temporally synchronized and geometrically calibrated data from an optical motion capture system, multi-baseline stereo cameras from multiple views, depth sensors, accelerometers and microphones. The dataset contains 11 actions performed by 7 male and 5 female subjects (in the range 23-30 years of age except for one elderly subject) 5 times. The total recording time is 82 minutes, which makes this dataset one of the biggest by the amount of video sequences it contains. The specified set of actions comprises of the following: (1) actions with movement in both upper and lower extremities, e.g., jumping in place, jumping jacks, throwing, etc., (2) actions with high dynamics in upper extremities, e.g., waving hands, clapping hands, etc. and (3) actions with high dynamics in lower extremities, e.g., sit down, stand up. Berkeley MHAD is popular [14], [6], probably due to the fact it allows to perform a multi-modal analysis of human motion and is easy to use.

TUM GAID: For depth based gait recognition and assessment, this challenging multimodal recognition database [24] was proposed in 2014. This database simultaneously contains RGB video, depth and audio. The database contain 305 individual gait captures, acquired in different weather conditions and in a different context, i.e: the person walks normally or is wearing a backpack or coating shoes (some persons performed all the actions and some only a subset). Other databases for depth gait analysis are available, however, TUM GAID is the most cited and currently used. It is the only database which allows for multi-modal gait recognition using video, depth and audio features along with different acquisition conditions.

Dataset	Action number	Person number	Calibration	Annotation	Total seq
MSR Action 3D	20	10	no	skeleton joints	567
MSR Gesture 3D	12	10	no	segmentation	336
MSR Daily Activity 3D	16	10	no	skeleton joints	320
Berkeley MHAD	11	12	yes	temporal synchronization	660
TUM GAID	3	305	no	metadata	3370

Table I: Popular 3D Video datasets and their characteristics

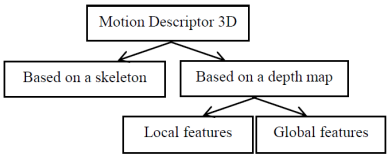


Figure 1: Motion descriptors classification

III. Motion descriptors for action recognition

Motion description is an essential part of human activity recognition. From a computational perspective, actions are best defined as four-dimensional patterns in space and in time. Methods which are able to discriminate the class of action being performed based on analysis of the video sequence combine a motion descriptor with a classifier. Most of the work on human action recognition published up to today relies on information extracted from 2D images and videos. However, with the availability of affordable depth sensors, this research area enlarged considerably with new studies dedicated to 3D. For a detailed review on action recognition, the works [25], [26] could be referred. These reviews, however, are not particularly oriented on RGBD-based methods as ours.

Motion descriptors can be categorized with respect to various criteria. A general classification feature-based scheme is shown in Figure 1. Further in this work we introduce 9 different motion descriptors for action recognition representing popular strategies and choices made by researchers in this field.

A very common approach for human action recognition is to track the human joints from the depth maps [27], [16]. A simple yet effective example of this approach is the depth-based algorithm by Xia et al. [11]. A view-invariant posture representation was devised using histograms of 3D joint locations (HOJ3D) within a modified spherical coordinate system. The positions of the joints in time form the 3D spatial histogram. HOJ3D were re-projected using LDA and clustered into k posture visual words. The temporal evolutions of these visual words were modeled by a discrete hidden Markov model (HMM).

Joint-based methods are popular, but will fail if the initial joints were estimated wrongly, which is still an issue. Moreover, if a very fine action is to be recognized (for example, a gesture), the joint-based methods lack precise information on shape and movement. For this reason, low-level attributes in depth images often outperform more high-level representations [15].

In 2010 Li et al.[12] proposed a new depth-map based method based on local features. They use an expandable graphical model to explicitly model the temporal dynamics of the actions and propose to use a bag of 3D points extracted from the depth map to model the postures. To select the points, they project the depth map onto the orthogonal Cartesian planes and further sample a specified number of points at equal distance along the contours of the projections. Selected points are then clustered in order to obtain salient postures. A Gaussian Mixture Model is used to globally model the postures by the distribution of points, and an action graph is constructed from the training samples to encode all actions that need to be recognized. This method gives better results than the 2D silhouette based action recognition, and the final descriptor is very compact. However, the cross-subject activity recognition results reported are low due to the fact that the proposed sampling scheme is view dependent. Secondly, the descriptors looses spatial context information between interest points, which could be a problem when using the method in a real life scenario.

Simplified motion-flow based descriptors could also be used [28], [13]. Munaro at et al. [29] proposed a global descriptor which takes the direction and magnitude of motion of every body part into account. For that, they first identify the human on the point cloud and then center a 3D grid around it. This grid divides the space around the person into a number of cubes. The flow information (direction and force) as a mean and a summary of motion vectors extracted from each cube is used as a motion description. Motion flow is calculated by using the KD-search algorithm and the color distance in HSV space. A single descriptor is concatenated for every video sequence. The published results are good, however, the descriptor is not view-independent and the task of the aligning the video frames in time is not addressed. Moreover, we can imagine the color information to be not of a great use when person tracked has solid color clothes making it hard to establish point correspondences based on color similarity.

Hadfield et al. [13] propose a novel local motion descriptor for RGBD video sequences. The descriptor encodes the 3D orientation of flow vectors around established interesting points extracted from evenly spaced regions. The nature of the local motion field is described using a spherical histogram in the velocity domain: the contribution of each flow vector to the histogram is weighted based on the magnitude of the flow vector. To remove a 3D rotation ambiguity and to make the descriptors completely consistent, the invariance to camera roll is encoded and the direction of the motions of flow vectors within the sub-region histograms is made rotationally-invariant. An interesting approach is also to perform PCA on the local region of the motion field, which finally leads to a descriptor which is invariant to all 3 types of camera viewpoint change, next to being robust to outlier motions. To obtain a global descriptor, the video sequence is divided into space-time blocks, each of which is encoded independently to provide the final description of the sequence. In this article it was proved that normalization and adaptation of the features so that they are scale and view point invariant improves the overall recognition of the system. However, the method is used for local interest points, discarding many relational information about the movement. Sequence time block-division schemes can also lead to wrong results in the recognition.

A more advanced HON4D (histogram of oriented 4D surface normals) descriptor [15] is analogous to the histogram of gradients in color sequences and extends the histogram of normals in static point clouds. In order to construct HON4D, the 4D space (XYZ, t) is initially quantized (in order to get the grid representation) using a regular 4D extension of a 2D polygon, namely, a 600-cell Polychoron. Then the normal to the surface in the 4D space represented as the set of points is computed and normalized by the length. To form a 120-dimensional motion descriptor, researchers compute the corresponding distribution of 4D surface normal orientation for each bin. Videos are divided in parts and final descriptor is a concatenation of individual parts descriptors. The results show that the motion descriptor outperforms several earlier descriptors [18], [16], however, they do not take into account the movement in the y and x direction, building their descriptor based on the change of depth.

A notion of hierarchy can be successfully employed in 3D motion descriptors. Kong et al. [30] improve the algorithm [15] using kernel descriptors alongside with the surface normals. They present a 3D gradient kernel descriptor which is a low-level depth sequence descriptor with the ability to capture detailed information by computing a pixel-level 3D gradient. The descriptor captures the change in shape of the 3D surface in time in the following way: Firstly, the 3D normals are computed and they are projected to a learned set of compact KPCA basis vectors [31]. The kernel is measures the similarity between orientations of the gradient for corresponding pixels from different patches of a video. A Bag-of-words method is used to build a hierarchical structure upon the low-level patch features to produce mid-level feature vectors. EMK [32] is employed over the output of the 3D kernel descriptor. The method achieves state-of-the-art performance for action recognition using depth data, but it is computationally expensive. It is also not explained how exactly the correspondences between pixels in different frames of the video are estimated.

Spatio-temporal features based descriptors gained lots of attention recently [14], [6]. Yang Xiao et al. [33] proposed a 3D trajectory shape descriptor for unconstrained RGBD video. To extract the 3D dense trajectory feature, the candidate feature points are densely sampled in each RGB frame and tracked using optical flow first. Then, by mapping the 2D positions of the RGB trajectory points to the depth map, motion information along the depth direction can be intuitively captured to form 3D trajectories. This method is a good illustration of how to enrich the 2D optical flow. To obtain the global representation of RGBD

videos, researchers combine several earlier approaches: Motion Boundary Histogram along the depth direction and trajectory shape descriptors [34] encoded using Fisher Vector [35]. It is however, not justified if the estimations of the 3D flow using the 2D and depth information gives good results in general.

Zhang et al. [14] propose discriminative and robust LST features named 4-D color-depth (CoDe4D) that incorporate both intensity and depth information acquired from RGBD cameras. The feature detector constructs a saliency map through applying independent filters in the XYZt dimension to represent texture, shape and pose variations, and selects its local maxima as interest points. A multichannel orientation histogram adaptive MCOH descriptor applies a 4-D support region, which is adaptive to linear perspective view changes, on each interest point. Then, image gradients of color-depth patches within the support region are computed and quantized using a spherical coordinate-based method to form a final feature vector. The method is interesting, however, the use of the color and texture information in the descriptor can be a disadvantage when a general real-case intra-person scenario application is aimed for. In this case it is proposed to tune the weighting parameters of the descriptor in order to weigh the depth information more and provide different weighting schemes for the datasets tested.

With the recent advances in human activity recognition, researchers are also addressed a challenging task of a group action recognition. Znang et al. [6] propose to use LST features which they call Adaptive Human-Centered (AdHuC) features. As in the previous method, their features are adapted to depth. To incorporate spatio-temporal and color-depth information in XYZt space researchers use a cascade of three filters: a pass-through filter to encode cues along the depth dimension, a Gaussian filter to encode cues in XY space, and a Gabor filter to encode time information. Then the color and depth cues are fused to form a saliency map. Local maximums on this map are then the LST features. The spatial-color-time based descriptor is then calculated for each point. The HOG3D [4] descriptor is modified in order to incorporate multichannel information. The final descriptor has a histogram form and concatenated of the per-channel descriptors.

IV. Motion descriptors for gesture recognition

Nowadays computer applications require new ways of interaction, especially within the growing Virtual Reality domain. For that reason, human-computer interaction and particularly, gesture recognition, became a very popular field of research in the last few years.

A gesture can be defined as a physical movement of the hands, arms, face and body with the intent to convey information or meaning [36]. Research in hand gesture recognition aims to design algorithms that can identify explicit human gestures. Gesture recognition dedicated motion descriptors are similar to the activity recognition ones with the difference that the descriptors should be capable of capturing very fine movements. Hand gestures are more difficult to recognize than body gestures due to the fact that the motions are more subtle, there are more degrees of freedom and serious occlusions occurs between the fingers. A detailed recent review on the advances in this area can be found in [37]. In this review, all the specifics of gesture recognition task are discussed along with the proposed algorithms. However, this paper tends to capture all the steps of gesture recognition process from detection up to classification and is not particularly dedicated to dynamic gesture recognition using 3D motion descriptors. The major part of the methods reviewed do not exploit the 3D point cloud based gesture recognition, which became particularly popular after 2012. For this reason, we also include in an overview of several motion descriptors applied to the gesture recognition on 3D data.

Similarly to full body motion description, many different types of visual features have been proposed for hand gestures. Early works uses 2D information and builds descriptors for the 2D silhouette of a hand. Due to the ambiguity of 2D data, the accuracy of such methods was not high. The latest dynamic gesture recognition methods use depth information and their 2D counterparts.

Model-based methods are very popular. Researchers often use the positions or the rotation angles of the joints from the skeleton structure of the fingers as the visual features [38]. An alternative approach is to use some form of geometric features extracted from a depth sequence [21]. For example, a shape silhouette can be used as a descriptor [39] or the cell occupancy information [40] similar to [17] can be used as feature. This results in approaches which are less dependent on separate segmentation and tracking algorithms.

Recently, 3D dynamic gesture recognition methods similar to action recognition ones are starting to avoid human body models and focus more on depth-based ones.

Cirujeda and Binefa [21] propose to use a Covariance matrix composed of the selected features from a 3D depth video sequence frame. Their descriptor doesn't use the absolute features themselves, but exploits representations of complex interactions between variations of 3D features in the spatial and temporal domain. It helps to make the descriptor robust to inter-subject and intra-class variations. The feature vector is the result of experimenting with several low-level cues and includes information about depth itself combined with other coarse observations such as first and second image derivatives, gradient magnitude, curvature and temporal information. The idea of their descriptor is to measure how several variables change together, capturing the intrinsic correlation between distributions of the involved cues. The final sequence descriptor is a concatenation of the three scene-wise covariances in its vectorized form. The descriptor captures the global motion patterns. The method is easily generalizable for action recognition and outperforms [20], [15]. It is also independent of the sequences length and from the cluttered background, however, it doesn't explicitly use the information about the movement (i.e. force and direction), which still can be useful in action recognition.

Descriptor	Year	Locality	Dimensionality	View-invariance	Accuracy, %	Complexity	Classifier	Dataset
HOJ3D [11]	2012	local	mid(≈ 1008)	yes	mid (78.97*)	low	HMM	MSR Action3D, custom (10 actions).
Bag of 3D points [12]	2010	local	low	no	low (74.7**)	low	NN	Custom, 20 actions, complex combinations.
3D grid-based descriptor[29]	2013	global	mid (≈ 5760)	no	mid (87.4**)	low	NN	IAS-Lab Action Dataset (15 actions, 12 person).
HON4D [15]	2013	global	low (≈ 120)	yes	high (88.89*)	low	SVM	MSR Action 3D, MSR Gesture 3D, MSR Actionv 3D Pairs, MSR Daily Activity 3D.
3D Flow descriptor[13]	2014	local	low(≈ 144)	yes	high(36.9**)	high	SVM	Hollywood 3D (14 actions, multi-cam setup).
3D kernel descriptor [30]	2015	hierachical	high (≈ 13824)	yes	high (92.73*)	high	SVM	MSR Action 3D, MSR Action 3D Pairs, and MSR Gesture 3D.
3D Trajectories [33]	2014	global	low (≈ 96)	no	mid (29.76**)	low	SVM	Hollywood 3D.
CoDe4D [14]	2016	local	high (≈ 21600)	yes	high (86**)	high	SVM	Berkeley MHAD, $ACT4^2$, MSR Daily Action 3D, UTK Action3-D.
AdHuC [6]	2015	local	low	yes	high (85.7*)	high	SVM	Berkeley MHAD and $ACT4^2$.

Table II: Motion descriptors for action recognition. **Reported accuracy is for MSR Action 3D dataset when available as reported by authors. **Accuracy reported for other dataset.*

Recently Ohn-Bar and Trivedi [1] proposed a combination of global low-level spatio-temporal features for gesture recognition in naturalistic driving settings. Their feature set is combining other earlier features: motion history image (MHI) [41] extension and HOG features. Several descriptors are tested along with different fusion schemes to establish the better-performing ones. Moreover, only compact descriptors are used, so the resulting one is small in dimensionality and fast to compute. For this work, depth and color data descriptors were extracted separately and their performance was compared. The best combination is Extended HOG2 + Extended MHI paired with DTM descriptor. This work shows that the approach to use a descriptor on color and depth data separately and fuse them in the final step of the algorithm works very well.

V. Motion descriptors for gait analysis

Lately, the subject of gait recognition and analysis from 3D data became popular. Gait is a manner of walking on a solid substrate. Observation of gait can provide early diagnostic clues for a number of movement disorders such as Parkinson's disease, cerebral palsy, stroke, arthritis, chronic obstructive pulmonary disease and many others. A general review on the subject of gait analysis is [42] and [43] overview all the recent advances of skeleton-based gait recognition. In our review, we provide the information on the use of motion descriptors in 3D gait assessment and recognition tasks.

Descriptors for gait recognition commonly include the biometrics parameters, because intra-person variability is no longer an issue as in the case of action recognition. Motion information is the part of information used to describe a gait pattern. For this reason usually the motion descriptors used for 3D gait recognition are more simple and compact. Despite the fact that RGBD cameras are popular for gait assessment tasks, it is quite common to use 2D projections in order to obtain a gait descriptor. Moreover, a vast majority of modern gait recognition and analysis methods perform a 3D-2D transformation of the depth sequence to form a final gait descriptor [44], [45], [46] or use 2D sensors directly [47]. The most well-known 2D gait descriptor is a Gait Energy Image [48], which is basically the average silhouette over one gait cycle. It was lately upgraded to a Gait Energy Volume (GEV)[49] by using information obtained from 3 Kinect sensors. GEV is derived by averaging all the voxel volumes over a gait cycle.

Similar to action and gesture recognition, 3D gait recognition methods can be categorized as methods based on skeleton joints [50], [51] (model-based) and methods based on depth images [52] (model-free). Skeleton-based methods are similar to analogue methods for action recognition: descriptors are based on the spatial and temporal position of the human skeleton joints or a human body model is used. Depth based gait features work on detailed information about shape and depth variation of a walking individual and do not require a model fitting.

Kwolek [44] propose a view independent motion-based algorithm for gait recognition using a multi-camera setup. They use particle swarm optimization for full-body motion tracking. A 3D human-body model is also proposed in order to improve the results. The final descriptor is a gait signature composed of the dynamic distances between joints projected to a 2D plan and evaluated through time of a single gait cycle. This approach is interesting because it uses the multi-camera setup in order to fit a 3D human model, however, the accuracy of the 3D model is limited due to the use of 2D images without depth information.

Tang et al. [45] introduce a 2.5D voxel gait model that includes only a one-side surface portion of the human body. A 2.5 gait model corresponding to a gait cycle is obtained from several Kinect depth frames. View-invariance is obtained by simply rotating the 2.5 gait model and synthesizing obtained views. The final descriptor is a color 2D image based on a combination of Gaussian and mean curvature [53] of the point cloud data. The method shows good results and avoids the high computational cost of 3D gait modeling, however, 2.5D gait model cannot address

the problems of the lack of robustness to covariates such as different appearance due to various clothes etc.

Lim et al. [52] propose real-time model-based gait tracking and analysis method using a depth image sensor installed on a robotic walker. The particle filter is adapted to the depth camera video sequences to obtain the spatio-temporal gait parameters. Segmented leg regions of the point cloud are also tracked using particle filtering improved by implementing a simple harmonic motion model. To simplify the problem, each particle represents the predicted leg model part. Spatio-temporal parameter data can be deduced from the tracked leg pose parameters. This methods proposes a computationally effective gait analysis method suited for clinical gait assessment, however, a specific setup is considered and the segmentation scheme proposed might not work in the case of a person with a movement disorder.

VI. Conclusion

This survey reveals the progress made in the last 6 years in the field of 3D motion descriptors for point cloud data. It is clear that 3D motion descriptors are developing towards more general and efficient descriptors, however, automatic motion analysis and classification remains an open problem. General applicable motion description algorithms are in trend. The latest approaches are compact, transformation invariant to the target object and robust to noise.

According to published results, approaches that model spatial and temporal statistics holistically for point cloud data show less promising results than LST feature points and projection based methods.

The main issue for many methods remains the time to extract the features from point cloud sequences. The features with the best recognition performance are often costly to compute. A detailed comparison of the time complexity of several popular descriptors can be found in [1].

Methods based on joint estimation usually provides compact and meaningful descriptors, and with the advances in skeleton joints recognition have great potential as well. Model-fitting methods remain popular for gait recognition and gesture recognition but for action recognition the main focus has shifted towards model-free methods.

Issues that must be addressed in future work in our opinion are: integration of all the cues for the better performance; computationally less expensive solutions; temporal alignment for the classification stage.

References

[1] E. Ohn-Bar and M. M. Trivedi, "A comparative study of color and depth features for hand gesture recognition in naturalistic driving settings," in *IEEE Intelligent Vehicles Symposium (IV)*, 2015, pp. 845–850.

[2] B. Steder, R. B. Rusu, K. Konolige, and W. Burgard, "Narf: 3d range image features for object recognition," in *Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, vol. 44, 2010.

[3] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 433–449, 1999.

[4] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference-BMVC*, 2008, pp. 275–1.

[5] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *ACM Proceedings of the 15th international conference on Multimedia*, 2007, pp. 357–360.

[6] H. Zhang, C. Reardon, C. Zhang, and L. E. Parker, "Adaptive human-centered representation for activity recognition of multiple individuals from 3d point cloud sequences," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1991–1998.

[7] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 1–4.

[8] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall, "A survey on human motion analysis from depth data," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Springer, 2013, pp. 149–187.

[9] N. Gelfand, N. J. Mitra, L. J. Guibas, and H. Pottmann, "Robust global registration," in *Symposium on geometry processing*, vol. 2, no. 3, 2005, p. 5.

[10] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta, "A discussion on the validation tests employed to compare human action recognition methods using the msr action3d dataset," *arXiv preprint arXiv:1407.7390*, 2014.

[11] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 20–27.

[12] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 9–14.

[13] S. Hadfield, K. Lebeda, and R. Bowden, "Natural action recognition using invariant 3d motion encoding," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 758–771.

[14] H. Zhang and E. Parker Lynne, "Code4d: color-depth local spatio-temporal features for human activity recognition from rgb-d videos," *IEEE Transaction on Circuits Syst Video Technol*, vol. 26, no. 3, pp. 541–555, 2016.

[15] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 716–723.

[16] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1290–1297.

[17] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, 2012, pp. 252–259.

[18] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 1057–1060.

[19] L. Liu and L. Shao, "Learning discriminative representations from rgb-d video data." in *IJCAI*, vol. 1, 2013, p. 3.

[20] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in

IEEE Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 2012, pp. 1975–1979.

[21] P. Cirujeda and X. Binefa, "4dcov: a nested covariance descriptor of spatio-temporal features for gesture recognition in depth sequences," in *2nd IEEE International Conference on 3D Vision*, vol. 1, 2014, pp. 657–664.

[22] C. Chen, B. Zhang, Z. Hou, J. Jiang, M. Liu, and Y. Yang, "Action recognition from depth sequences using weighted fusion of 2d and 3d auto-correlation of gradients features," *Multimedia Tools and Applications*, pp. 1–19, 2016.

[23] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *IEEE Workshop on Applications of Computer Vision (WACV)*, 2013, pp. 53–60.

[24] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll, "The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 195–206, 2014.

[25] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[26] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[27] J. Zheng and Z. Jiang, "Learning view-invariant sparse representations for cross-view action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3176–3183.

[28] S. Hadfield and R. Bowden, "Kinecting the dots: Particle based scene flow from depth sensors," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2290–2295.

[29] M. Munaro, S. Michieletto, and E. Menegatti, "An evaluation of 3d motion flow and 3d pose estimation for human action recognition," in *RSS Workshops: RGB-D: Advanced Reasoning with Depth Cameras*, 2013.

[30] Y. Kong, B. Satarboroujeni, and Y. Fu, "Hierarchical 3d kernel descriptors for action recognition using depth sequences," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1, 2015, pp. 1–6.

[31] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[32] L. Bo and C. Sminchisescu, "Efficient match kernel between sets of features for visual recognition," in *Advances in neural information processing systems*, 2009, pp. 135–143.

[33] Y. Xiao, G. Zhao, J. Yuan, and D. Thalmann, "Activity recognition in unconstrained rgb-d video using 3d trajectories," in *ACM SIGGRAPH Asia Autonomous Virtual Humans and Social Robot for Telepresence*, 2014, p. 4.

[34] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[35] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Computer Vision–ECCV*. Springer, 2010, pp. 143–156.

[36] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, 2007.

[37] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2015.

[38] S. Celebi, A. S. Aydin, T. T. Temiz, and T. Arici, "Gesture recognition using skeleton data with weighted dynamic time warping." in *VISAPP (1)*, 2013, pp. 620–625.

[39] Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 1093–1096.

[40] P. Suryanarayan, A. Subramanian, and D. Mandalapu, "Dynamic hand pose recognition using depth data," in *20th IEEE International Conference on Pattern Recognition (ICPR)*, 2010, pp. 3105–3108.

[41] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[42] M. J. Nordin and A. Saadoon, "A survey of gait recognition based on skeleton mode l for human identification," *Research Journal of Applied Sciences, Engineering and Technology*, 2016.

[43] A. Muro-de-la Herran, B. Garcia-Zapirain, and A. Mendez-Zorrilla, "Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications," *Sensors*, vol. 14, no. 2, pp. 3362–3394, 2014.

[44] B. Kwolek, T. Krzeszowski, A. Michalczuk, and H. Josinski, "3d gait recognition using spatio-temporal motion descriptors," in *Intelligent Information and Database Systems*. Springer, 2014, pp. 595–604.

[45] J. Tang, J. Luo, T. Tjahjadi, and Y. Gao, "2.5 d multi-view gait recognition based on point cloud registration," *Sensors*, vol. 14, no. 4, pp. 6124–6143, 2014.

[46] M. Hofmann, S. Bachmann, and G. Rigoll, "2.5 d gait biometrics using the depth gradient histogram energy image," in *IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems*, 2012, pp. 399–403.

[47] M. Alotaibi and A. Mahmood, "Automatic real time gait recognition based on spatiotemporal templates," in *IEEE Systems, Applications and Technology Conference (LISAT)*, 2015, pp. 1–5.

[48] J. Man and B. Bhanu, "Individual recognition using gait energy image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[49] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes, "Gait energy volumes and frontal gait recognition using depth images," in *IEEE International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–6.

[50] P. Chattopadhyay, S. Sural, and J. Mukherjee, "Frontal gait recognition from incomplete sequences using rgb-d camera," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 11, pp. 1843–1856, 2014.

[51] M. Milovanovic, M. Minovic, and D. Starcevic, "Walking in colors: human gait recognition using kinect and cbir," *IEEE MultiMedia*, vol. 20, no. 4, pp. 28–36, 2013.

[52] C. D. Lim, C.-Y. Cheng, C.-M. Wang, Y. Chao, and L.-C. Fu, "Depth image based gait tracking and analysis via robotic walker," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 5916–5921.

[53] P. Tosranon, A. Sanpanich, C. Bunluechokchai, and C. Pintavirooj, "Gaussian curvature-based geometric invariance," in *IEEE. 6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, vol. 2, 2009, pp. 1124–1127.

Automatic Processing Scheme for Low Laser Invasiveness Electro Optical Frequency Mapping mode

A.Boscaro[1,2], S.Jacquir[1], K.Sanchez[2], H.Terada[4], P.Perdu[2,3] and S.Binczak[1]
[1]Le2i UMR CNRS 6306, Univ. Bourgogne Franche-Comté, 9 avenue Alain Savary, 21078 Dijon, France
[2]Centre National d'Études Spatiales, 18 avenue Avenue Edouard Belin, 31401 Toulouse, France
[3]Temasek Laboratories, Nanyang Technological University, 50 Nanyang Drive, Singapore
[4]HAMAMATSU Photonics K.K., Japan
Phone: +33(0)770023076 Email: anthony.boscaro@u-bourgogne.fr

***Abstract*—Electro optical techniques are efficient backside contactless techniques usually used for design debug and defect location in modern VLSI. Unfortunately, the signal to noise ratio is quite low and depends on laser power with potential device stress due to long acquisition time or high laser power, especially in up to date technologies. Under these conditions, to maintain a good signal or image quality, specific signal or image processing techniques can be implemented. In this paper, we proposed a new spatial filtering by stationary wavelets and contrast enhancement which allows the use of low laser power and short acquisition time in image mode.**

***Index Terms*—EOP, EOFM, Stationary Wavelet Transform, Filtering, Contrast enhancement.**

I. Introduction and problem statement

VLSI failure analysis is facing endless renewed challenges induced by technology evolution. It obviously concerns fault isolation and defect location techniques. In addition to light emission techniques, methods based on laser exploit optical stimulation or optical properties of reflected beam [1]. Since its introduction, electro optical probing has become an established timing-analysis laser based technique [2], using Franz-Keldysh effect end free carrier absorption. In other words, these techniques use the observation of the emitted or reflected beam from the device under test (DUT). In the case of these optical techniques, it exists two modes: Point mode (Probing on one node) such as Electro-optical probing (EOP) or Laser Voltage Probing (LVP), or image mode Electro Optical Frequency Mapping (EOFM) also known as Laser Voltage Imaging (LVI) [3]. In our study, we will only focus on the second one. Nowadays with new VLSI technologies, 28 nm for instance, techniques using laser mode could be invasive. Consequently, it is not possible to stay several seconds on each pixel. Another problem resulting from image mode, is the noise which is the result of various phenomena especially the interaction between the light and the silicon, see Fig. 2a. To have an image with good quality in EOFM, it is possible to increase the laser's power but this solution could stress the integrated circuit (IC) or destroy it. In order to address that concern, a processing based on wavelets filtering and contrast enhancement has been suggested in this paper. The aim of this image processing scheme is to only modify the EOFM image because modification of the setup could be very complicated and expensive. This process allows experts to use a low laser to visualize the Regions of interest (ROI). Others image processing methods have already been implemented in the failure analysis (FA) community, especially for Time Resolved Imaging (TRI) [4], [5]. In the next section, EOFM background is described with an exemple of result. Then, the mathematical theory of the process is detailed. The third section is dedicated to applications. Examples with In-phase/Quadrature (IQ) images and different laser powers (10%, 20%, 40% and 100%) will be shown and compared in the discussion part. Finally, the conclusions are given with the potential perspectives.

II. Solution: filtering by SWT and contrast enhancement

A. Acquisition setup

FA techniques which take advantage of the laser properties have proved their effectiveness. Two main modes can be distinguished: point mode (EOP, LVP) and image mode (EOFM, LVI). In this study, only the image mode is taken into account. For more details about EOFM, reference [3] explains the physical principle. Applications such as EOFM, only require the identification of the core of a spot. There is no need to precisly label spot's edges, but identifying all frequencies of signals located by the spots is a key factor for the rest of the analysis. It allows the engineer to know on which node he has to probe. An example of this result is illustrated in Fig. 2a. The image is relatively noisy and some

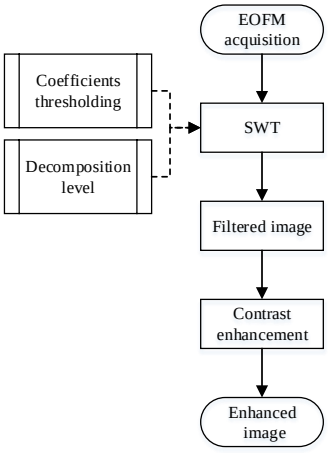
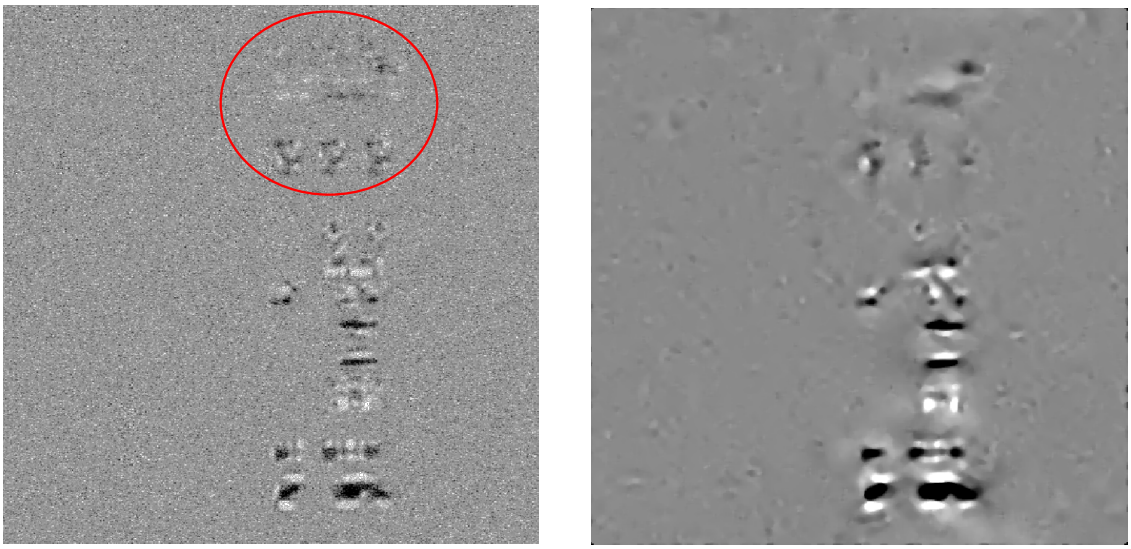


Fig. 1: Flowchart of the process.



(a) (b)

Fig. 2: Results on EOFM IQ images with 100% without image processing (a) and 10% (b) of laser power with new post processing .

areas are not correctly visible even with 100% of laser power. Our new process allows one to obtain the kind of result shown in Fig. 2b. The flowchart' steps are illustrated below in Fig. 1 and explained in next sections.

B. Filtering by Stationary Wavelet Transform (SWT)

As mentionned previously, scanning a device in advanced technology can be invasive event even with quite low laser power or long laser exposure time. It is more invasive for recent VLSI technologies. To overcome this problem we have introduced a spatial filtering based on a combination of image processing steps. Now, each step will be detailed. The fisrt step is to remove the noise without degrading the image. Several filtering methods in image processing [6] exist, but here, a filtering process which will not introduces new pixel values is recommended in order to keep the useful information. With this purpose in mind, the filtering by stationary wavelets has been extensively studied and presented in the literature [7]. It is generally better than linear filtering such as mean, gaussian filters and non-linear filtering as median and Wiener [8]. It is recalled that in failure analysis, wavelets have already been applied to improve the signal to noise ration (SNR) in EOP [9], [10]. In the process, the first step is to decompose the image with the Stationary Wavelet Transform (SWT) at a specific level. It is similar to the Discrete Wavelet Transform (DWT) except the signal is never sub-sampled and instead the filters are up sampled at each level of decomposition. The SWT is an inherent redundant scheme, as each set of coefficients contains the same number of samples as the input. So for a decomposition of N levels, there is a redundancy of 2N (N is the signal's length). Once the image is decomposed, we obtain differents kinds of coefficients: Horizontal, vertical and diagonal [11]–[13]. Concerning the choice of the decomposition level, it will be discussed later in this paper. After that we apply a threshold to each of them in order to keep the useful coefficients. According to the literature, differents kinds of thresholding exist for wavelets coefficients [8]:

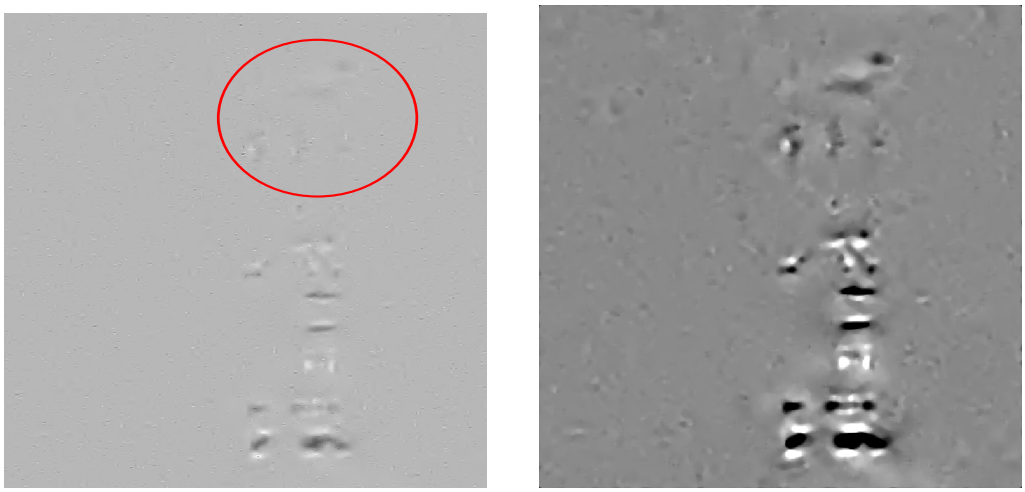
- ***Soft thresholding*** : The absolute value of all the wavelets coefficients are compared to a threshold T. If this value is greater than T the threshold is subtracted from any coefficient that is greater than the threshold. Others are set to zero.
- ***Hard thresholding*** : Hard thresholding sets any coefficient less than or equal to the threshold T to zero. Others are preserved.
- ***Universal thresholding*** : the value called *universal*, is defined by :

$$T_{thresh} = \sigma\sqrt{2log(N)} \tag{1}$$

where N is the length of the image and σ the noise's standard deviation. Here, T_{thresh} is used with hard thresholding. In several applications, noise is most of the time white and gaussian. This kind of white guaussian noise (WGN) is a random signal with constant power spectral density. WGN whose representation is given by (6), is independant and identically distributed (i.d.d) and drawn from a zero-mean normal distribution with variance σ^2.

$$WGN \sim N(0,\sigma^2) \tag{2}$$

According to (2), only σ^2 is unknown. That is why wavelets are useful in our study. Here, only the standard deviation could be estimated. In rare cases the noise is assumed but in others, it can be estimated by using the Median Absolute Deviation (MAD). This method has been introduced by Donoho and Johnstone in 1994 [14]. MAD is the median absolute deviation of the empirical wavelet coefficients corresponding to the first level $j1$. The reason for using these first level coefficients for the variance estimation, is that they are mostly constitued of



(a) (b)

Fig. 3: Contrast enhancement on filtered EOFM IQ images with 10% of laser power. (a) Filtered image, (b) Enhanced image.

noise [15]. By consequences the estimated variance is given by

$$\sigma^2 = \left(\frac{MAD}{0.6745}\right)^2 \tag{3}$$

where 0.6745 is the 0.75- quantile of the standard normal distribution and

$$MAD((w_j)) = Median((|w_j|)_j) \tag{4}$$

with w_j, the wavelets coefficients. Thus, σ^2 is now known and the threshold given in (5) can be computed. Finally, the inverse transform is applied to recover the denoised image. Then, some of the ROIs could be partially or not visible. That is why we use constrast enhancement in order to make them more readable.

C. Contrast enhancement

In some applications, it is difficult to see all the characteristics of the image. That is reason why image enhancement is used to solve this problem. Contrast enhancement is one of the commonly used image processing methods [16]. In this study the contrast is enhancend by mapping the values of the input intensity image to new values according the following rule: 1% of the image is saturated at low and high intensities of the input image. An example of image improvement is illustrated in Fig. 3 with IQ images.

III. Application, results and discussion

A. Examples of application

Here the process is applied on EOFM IQ images with (10%, 20%, 40% and 100%) of the laser's power. Results are reported in Fig.(4-7). The acquisition focuses on some areas in the right corner of the image which are not visible at all with low power. The outcome of ROI identification by SWT filtering is given in Fig. 2b. For this step, SWT was used with nine decomposition levels and Daubechies wavelet. For the thresholding step, the noise is evaluated automatically by estimating its standard deviation σ. The most emissive ROI are detected but others are not perfectly viewable, see Fig. 7b red circle. The next step of the process is to highlight hidden aeras with contrast enhancement. Once the contrast is improved, the following image is obtained and illustrated in Fig. 7c. Finally, hidden ROIs are now visible with only 10% of laser power.

B. Discussion

This part addresses a discussion concercing the prior parameters for completly automate the process. More precisly this section deals with the choice of the decomposition level to justify the choice of a filtering by SWT with low laser power. Here, a qualitative approach is used by using the structural similarity index measurement (SSIM) [17] and a reference image (ground truth). This index measures the image quality in terms of luminance, contrast and structure by comparing with a reference image. It gives a score between 0 and 1.

1) Decomposition level: As presented in [9], in the wavelets theory, the maximum decomposition level (DL) is given by the following equation

$$DL_{max} = log_2 N, \tag{5}$$

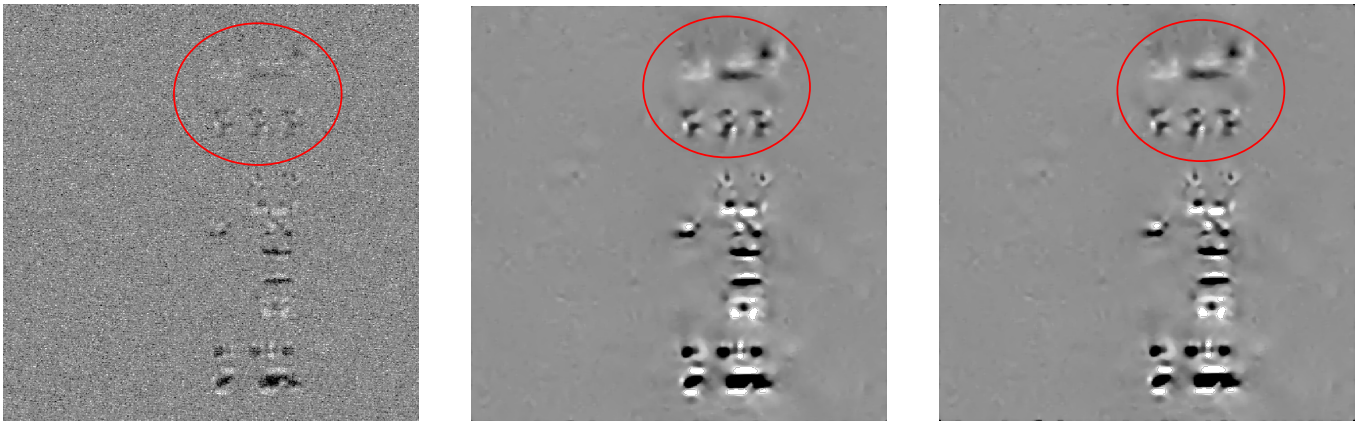
with N the image' size. For example, if the used image has a size of 512×512 pixels, the DL_{max} is equal to $log_2(512) = 9$. On the one hand, DL_{max} can be computed easily. On the other hand, the expert can choose manually the DL. To show the impact of the DL, SSIM is computed as a function of the DL (from 1 to 9) for different laser powers. The results are illustrated in Fig. 8. We can notice that for high DL, the SSIM is the best for all tested laser powers. The most important aspect in this part is that results are sensibly the same for low and high laser power. Results prove the efficiency of our process in terms of image reconstitution. Engineers can have a gain between 5 and 10 about the laser power. Thus the invasive effect is considerably reduced.

(a) (b) (c)

Fig. 4: Results on EOFM IQ images with 100% of laser power. (a) Original image / (b) Image denoised with stationary wavelets transform / (c) Image with contrast enhancement.

(a) (b) (c)

Fig. 5: Results on EOFM IQ images with 40% of laser power. (a) Original image / (b) Image denoised with stationary wavelets transform / (c) Image with contrast enhancement.



(a) (b) (c)

Fig. 6: Results on EOFM IQ images with 20% of laser power. (a) Original image / (b) Image denoised with stationary wavelets transform / (c) Image with contrast enhancement.

(a) (b) (c)

Fig. 7: Results on EOFM IQ images with 10% of laser power. (a) Original image / (b) Image denoised with stationary wavelets transform / (c) Image with contrast enhancement.

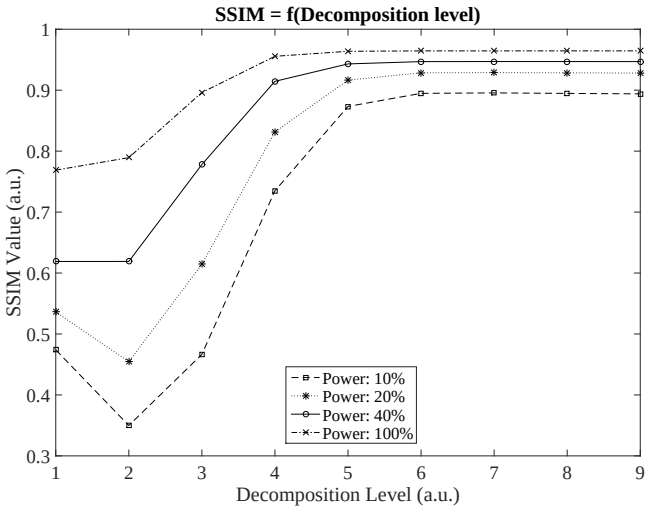


Fig. 8: SSIM index as a function of the Decomposition Level for different laser powers (10%, 20%, 40% and 100%).

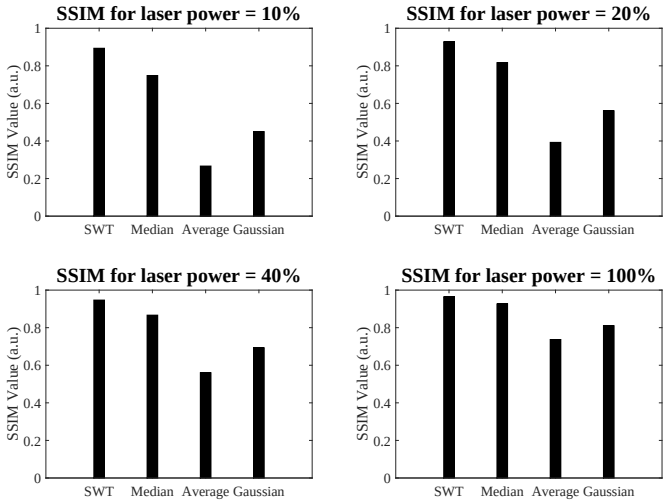


Fig. 9: SSIM index as a function of different kinds of filters (SWT, Median, Average, Gaussian) for different laser powers.

2) Choice of SWT filtering: In order to prove the efficiency of the SWT filtering, this filter is compared to others with a quantitative approach based on structural similarity index measurements SSIM. For different laser powers and different kinds of filters, the SSIM is computed. Results are given by Fig. 9. On this bar diagram, the SSIM value is always higher with SWT filtering. Also notice in the previous part, results are still the same for 100% and 10% of laser power.

IV. Conclusion and Perspectives

With new technologies like 28 nm and 15 nm, using high laser power is a real problem. But selecting low laser power provides very noisy images and engineers do not know where the ROIs appear. Filtering a noisy image is not a trivial task, especially in EOFM where detected areas are not always visible. In this paper, a filtering method combining different image processing steps has been reported. This process could be useful for the FA community because it allows to significantly reduce the invasive effect of the laser and thus saving time for the expertise. The process is fully automated and in the event of the engineer is not satisisfied with the result, he can manually adjust all the parameters. Furthermore, this study could be a complement to others existing images enhancement methods [18], [19].

In terms of perspectives, this process allows the FA community to use signal processing tools to solve their difficulties in terms of noisy acquisitions.

V. Acknowledgments

This work has been supported by Regional Council of Burgundy (Dijon, France) and the French National Spatial Center. Authors would like to especially thank Hamamatsu Photonics for its technical support (TriPHEMOS).

References

[1] Chin Jiann Min, Narang Vinod, Zhao Xiaole, Tay Meng Yeow, Phoa Angeline,Ravikumar Venkat, Ei Lwin Hnin, Lim Soon Huat, Teo Chea Wei, Zulkifli Syahirah and others, "Fault isolation in semiconductor product, process, physical and package failure analysis: Importance and overview", *Microelectronics Reliability* ,vol. 51, pp. 1440–1448, 2011.

[2] Kindereit U, "Fundamentals and Future Applications of Laser Voltage Probing", *IEEE-IRPS* , pp. 162–172, 2014.

[3] Perdu, P., Bascoul, G., Chef, S., Celi, G. and Sanchez, K, "Optical probing (EOFM/TRI): a large set of complementary applications for ultimate VLSI", *IPhysical and Failure Analysis of Integrated Circuits (IPFA)*, pp. 119–126, 2013.

[4] Boscaro, A., Chef, S., Jacquir, S., Sanchez, K.,Perdu, P., and Binczak, S, "Automatic emission spots identification in static and dynamic imaging by research of local maxima", *ISTFA 2014: Conference Proceedings from the 40th International Symposium for Testing and Failure Analysis, Houston, Texas, USA*, pp. 322–326, 2014.

[5] Chef, S., Jacquir, S., Sanchez, K.,Perdu, P., and Binczak, S, "Filtering and emission area identification in the Time Resolved Imaging data", *ISTFA 2012: Conference Proceedings from the 38th International Symposium for Testing and Failure Analysis, Phoenix, Arizona, USA*, pp. 264–272, 2012.

[6] Burt, PJ., "Fast filter transform for image processing", *Computer graphics and image processing* ,vol. 1, pp. 20–51, 1981.

[7] Nason, GP., and Silverman, BW., "The stationary wavelet transform and some statistical applications", *Lecture Notes in Statistics-New York-Sringer Verlag-*, pp. 281, 1995.

[8] Walker, JS., "Tree-adapted wavelet shrinkage", *Advances in Imaging and Electron Physics*,vol. 124, pp. 343–394, 2002.

[9] Boscaro, A, Jacquir, S, Sanchez, K, Perdu, P, Binczak, S, "Improvement of signal to noise ratio in electro optical probing technique by wavelets filtering", *Microelectronics Reliability*,vol. 55, pp. 1585–1591, 2015.

[10] Chef, S, Jacquir, S, Sanchez, K, Perdu, P, Binczak, S, "Frequency mapping in dynamic light emission with wavelet transform", *Microelectronics Reliability*,vol. 53, pp. 1387–1392, 2013.

[11] Wang, XH and Istepanian, Robert SH and Song, Yong Hua, "Microarray image enhancement by denoising using stationary wavelet transform", *NanoBioscience, IEEE Transactions on*,vol. 2, pp. 184–189, 2003.

[12] Osman, AM "Enhancement of Photon and Neutron Images by Denoising Using Stationary Wavelet Transform", *Armenian Journal of Physics*,vol. 4, pp. 154–164, 2011.

[13] Naga, R and Chandralingam, S and Anjaneyulu, T and Satyanarayana, K "Denoising EOG signal using stationary wavelet transform", *Measurement Science Review*,vol. 12, pp. 46–51, 2012.

[14] D.L. Donoho and J.M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage", *Biometrika*, vol. 81, pp. 425–455, 1994.

[15] W. Hrdle, G. Kerkyacharian, D. Picard and A. Tsybakov, "Wavelet thresholding and adapation", *Wavelet, Approximation, and Statistical Application*,pp. 193–213, 1998.

[16] Kotkar, Vijay A and Gharde, Sanjay S, "Review of various image contrast enhancement techniques", *International Journal of Innovative Research in Science, Engineering and Technology*,pp. 2786–2793, 2013.

[17] Wang, Zhou and Bovik, Alan Conrad and Sheikh, Hamid Rahim and Simoncelli, Eero P, "Image quality assessment: from error visibility to structural similarity", *Image Processing, IEEE Transactions on*,pp. 600–612, 2004.

[18] Chef, S, Jacquir, S, Sanchez, K, Perdu, P, Binczak, S, "Pattern image enhancement by extended depth of field", *Microelectronics Reliability*,vol. 54, pp. 2099–2104, 2014.

[19] Chef, S, Jacquir, S, Sanchez, K, Perdu, P, Binczak, S, "Spatial correction in dynamic photon emission by affine transformation matrix estimation", *Physical and Failure Analysis of Integrated Circuits (IPFA), 2014 IEEE 21st International Symposium on the*, pp. 118–122, 2014.

NoC based virtualized FPGA as cloud Services

Hiliwi Leake Kidane, El-Bay Bourennane
Laboratorie Le2i
Université Bourgogne Franche-Comté
21000 Dijon, France
Email:hiliwi-leake.kidane@u-bourgogne.fr, ebourenn@u-bourgogne.fr

***Abstract*—Web-based applications are increasingly demanding many computationally intensive services. On the other hand, FPGA-based hardware accelerators(HwAcc) provide good performance in accelerating computationally intensive applications. In addition, some FPGAs support a dynamic partial reconfiguration (DPR) techniques to virtualize and share the FPGA underlying hardware resources in time multiplexing during run-time to save resource and power consumption. Integrating FPGA in a cloud environment is an indispensable way to improve efficiency and provide acceleration services to demanding users. More importantly, in recent years it was proved that FPGA resources deployed in a cloud environment can be accessed with the same OpenStack software technology used to access virtual machines. However, the performance of the virtualized FPGA is highly dependent on the communication medium used to interconnect the virtualized FPGA resources and the control manager. After analyzing the possible interconnect mediums, we have selected Network-on-Chip (NoC) which support parallel communication as the efficient medium for accelerators. Consequently , we propose a NoC based virtualized FPGA as cloud Services. Two virtualized FPGA-based cloud service: Hardware Accelerator as a Service(HAaaS) and Reconfigurable Region as a Service(RRaaS) are proposed in this paper. The NoC provides layered and parallel communication between the virtualized regions of the FPGA and helps them to communicate their status and exchange data through the routers connected to them. A 2x2-mesh NoC based reconfigurable accelerators for image analysis and matrix computation are implemented and tested showing a promising result for more scalable systems in cloud computing.**

***Index Terms*—Cloud Computing; Virtualized FPGA; Hardware accelerators; Network-on-Chip;**

I. Introduction

Web based applications are increasingly demanding many services which are highly resource consuming that can not be performed locally [1]. Cloud computing minimizes these problems by offering software, platform and infrastructure as a service. The cloud makes it possible for a user to access information from anywhere at any time [2]. It uses billing mechanisms to use these resources on the basis of their consumption, allowing on-demand model: pay-per-use [3].

On the other hand, Field programmable Gate Array (FPGA)-based hardware accelerators(HwAcc) provide good performance in accelerating computationally intensive applications [4] like multimedia image analysis. In addition, some FPGAs support a dynamic partial reconfiguration (DPR) techniques to virtualize and share the FPGA underlying hardware resources in time multiplexing during run-time to save resource and power consumption.

Consequently, cloud provider can improve their computing performance and provide accelerating service by integrating virtualized FPGA in a cloud environment. More importantly, in recent years, it has been demonstrated that FPGA resources deployed in a cloud environment can be accessed with the same Openstack software [5] technology used to access virtual machines [6], [7], [8]. OpenStack is a cloud operating system that controls large pools of compute, storage, and networking resources throughout a datacenter [9].

However, communication issues could hamper the proper operation of such method if proper communication medium is not proposed. When the number of processors in a given FPGAs increases, the point to point connection using the ordinary wire or bus based communication is not efficient and reliable. On the other hand, the Networks-on-Chip (NoCs) have recently emerged as a promising concept to support communication on Multi-Processor SoCs [10] due to their scalable and layered architecture. Thus, in this paper we propose a method that takes best of both worlds: the improved communication features of NoCs with the adaptability on demand provided by dynamically reconfigurable systems in order to integrate the virtualized FPGA and provide its resources as a services.

II. Background

Before proceeding to the main contributions of this paper, we will first provide some basic concepts and terminology.

A. Cloud Computing

Cloud computing is the provision of computing resource from remote on-demand [16]. The cloud services are mainly divide into three categories [2]: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) as shown in Fig. 1.

There are some essential characteristics to consider during any cloud service proposal [2] [16] [3]. The first characteristic is on-demand self service or service without human interaction. The second characteristic refers to the availability of the service over the network and accessibility by standard client platforms. The third characteristic is related to the resource pooling or parallelism. In other words, the service must serve multiple clients at the same time. The other important characteristic is rapid elasticity of the service. When clients want to extend the resource they are using, the service should allow them to scale it dynamically. The final and most

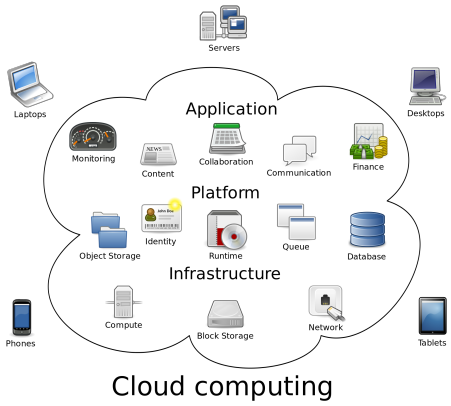


Fig. 1. Cloud Computing (source: wikipedia)

important characteristic is the service has to be measurable and the control has to measure and report service usage for payment.

B. Network-on-Chip (NoC)

1) building blocks: All NoCs have three fundamental building blocks, namely, switches (also called routers), Network Interfaces (NIs) and links [12] as shown in Fig. 2. A Router is responsible for the routing or directing of data based on the protocol defined at that moment. The router contains an arbiter, a buffer and an input-output to connect ports. Links are the channels that connect router to router or router to NI. NI is an intermediate between the router and the processing element (PE) connected to the router. The network interface (NI) is responsible for packetization and depacketization of data traffic, in addition to the conventional interfacing [13].

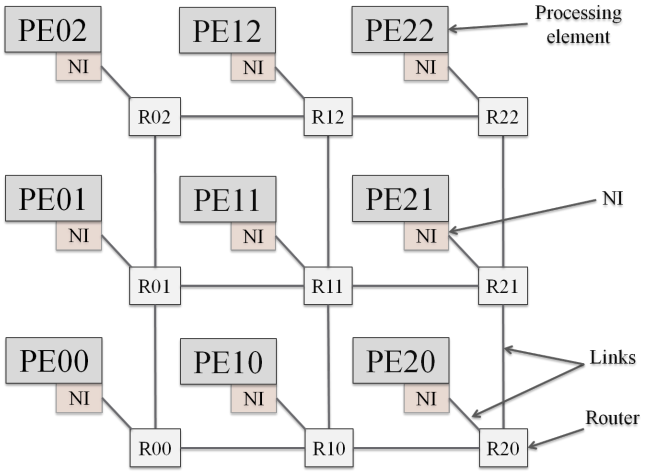


Fig. 2. Network-on-Chip (NoC) based communications of PEs

2) Topology: refers to the interconnection between NoC building blocks. The topology can be categorized as regular and irregular based on the distribution of the routers in the network. The regular topology also contains different varieties like mesh, mesh tours, ring, fat tree. The performance of NoC based communication strongly correlates with the topology selected for implementation [14].

3) Switching: determines how the data is transferring from one node to the other. The two common switching techniques are packet switching and circuit switching. Circuit switching, the link is used in a spatial division approach until the data transmission is completed. Whereas in the packet switching the links are used in a time division approached. Different packets of the same data can follow different path [15].

4) Routing: refers to the algorithm in the routers which selects an output port for the packet coming through its input port. There are different varieties of routing algorithms which give different results [13].

5) Flow Control: determines how network resources, such as channel bandwidth, buffer capacity, and control state, are allocated to a packet traversing the network [13]. An efficient design of flow control is crucial in order to get good performances.

C. Dynamic Partial Reconfiguration

Dynamic partial reconfiguration (DPR) is a technique that enables to dynamically modify preselected area of the FPGA at run-time and on demand. The DPR is not supported by all FPGAs. The Atmel FPGA and Xilinx FPGAs are some of the few FPGAs in the market allowing DPR. Figure 3 shows the basic premise of Partial Reconfiguration (PR). The reconfigurable region represented by ”A” can be used to configure and run different modes of ”A” in time multiplexing.

The general PR flow as stated in the Xilinx PR flow [11] is as follows. The logic in the FPGA design is divided into static logic and reconfigurable logics/modules. The static logic does not change its functionality during reconfiguration and it contains an embedded processor, an internal configuration access port (ICAP), DSP and other circuitry. The embedded processor runs the configuration management software which controls system transition from one configuration to another depending upon the adaptation conditions set by the application. The ICAP is used to load/transmit the partial bitstreams into the FPGA configuration infrastructure.

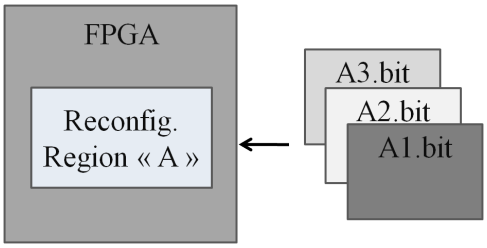


Fig. 3. Basics of Partial reconfiguration

The DPR process starts by implementing the static or top level HDL module which includes a black box for the partially reconfigurable modules and then synthesized to generate the top level netlist. The HDL of the different modes of the partially reconfigurable modules should be implemented

and synthesized separately to generate independent netlists. Subsequently, both the top level and partial reconfigurable module Netlists are imported to the PlanAhead to floorplan the placement and mapping of the Reconfigurable regions and reconfigurable modules. Placement restriction must be associated to reconfigurable IPs during DPR design to define their size, shape and position.

III. Previous Works

The use of virualized FPGA accelerator in datacenters for better performance and flexibility is presented in [17], [18], [19]. Authors in [20] presented a prototype for integrating virtualized FPGA accelerators in the cloud using partial reconfiguration and virtual communication. Similarly, [21] proposes a resource virtualization solution based on run-time partial reconfiguration to share reconfigurable resources among the underutilized microprocessors. In addition, [6], [7], [8] presented that FPGA resources deployed in a cloud environment can be accessed with the same Openstack software technology used to access virtual machines.

Bharathi and Neelamegam [1] propose a Reconfigurable Framework for Cloud Computing Architecture with three layer called service usage, service provider and service developer. If the demanding service from a user is not available, service usage module requests service developer module through service provider module to develop a new or modify existing service to satisfy the customer application need on demand. Reconfiguration takes place if the service is not available in the hardware. Reconfiguration bitstreams are generated in service developer module and transferred to the hardware with CPU control.

Knodel and Spallek [22] present a cloud service models and cloud hypervisor called RC3E, which integrates virtualized-FPGA based hardware accelerators into a cloud environment. The authors defined three types of service called RSaas, RAaaS and BAaaS. The RSaas service provides full access to reconfigurable resources for the user. The user can allocate a complete physical FPGA with own implemented hardware. In RAaaS, the FPGA is used as a simple accelerator and only the vFPGAs or virtual reconfigurable regions of different size are visible and accessible by the user. In the last service, BAaaS, only available applications and services are visible to the user. These applications and service use the virtual reconfigurable regions of the FPGA(vFPGA) in the background to accelerate specific applications.

Even though most of the authors provide detailed information about the advantages of integration virtualized FPGA for cloud computing, they all failed to give attention to the complexity of communication requirement for data as well as control signal communication.

IV. NoC based virtualized FPGA as cloud Services

In this paper, we have proposed two NoC based virtualized FPGA as service models for cloud computing called Hardware Accelerator as a Service(HAaaS) and Reconfigurable Regions as a Service(RRaaS). The NoC will help to easily access the reconfigurable IPs and send signal to the control up on completion of tasks.

The proposed virtualized FPGA (vFPGAs) as cloud services are based on 2D-mesh Hermes NoC [24] architecture. Each vFPGAs is connected to the nearest router through network interface(NI). The internal structure of the NI depends on the ports of the vFPGAs connected to it. The NI will also serve as an isolation during reconfiguration of the vFPGAs connected to it.

The architecture has three layers: the virualized FPGA or reconfigurable regions, the hypervisor which contains the static region of the FPGA and the application layer which gives interface to the external user as shown in figure 4. The Hypervisor is responsible for the management of reconfiguration and resources and many virualized machines are installed over a single hypervisor.

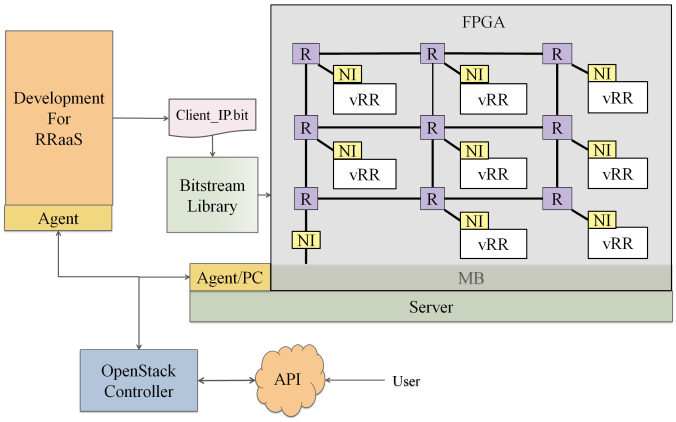


Fig. 4. Architecture of the NoC based virtualized FPGA as cloud Services

A. Hardware Accelerator as a Service(HAaaS)

In this service model, the user does not have direct access to the virtualized FPGA. When the user requests any accelerator, the intermediate control manager checks availability of the requested accelerator and if it exists, it establishes connection between accelerator and the user in the host machine. If not, the control manager demands to reconfigure the requested accelerator from the existing bitstream library as stated in the flow diagram in Figure 5.

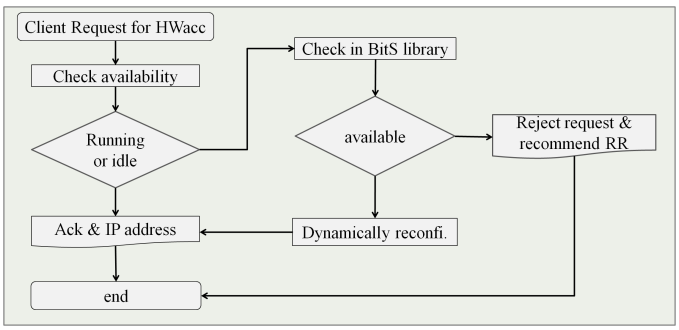


Fig. 5. Flow diagram of HAaaS

In case, the requested accelerator does not exist in the library at all, the control manager has to reject the request and recommends the user to use the other service RRaaS.

B. Reconfigurable Regions as a Service(RRaaS)

This services can be divided into two stages as development and provision. First, the user get access to select one of the available virtualized RRs based on their top-level architecture and capacity in terms of resourses. Next, the control manager send back the HDL template file of the selected vRR to the user to implement the detail design. Later, the user will send back the completed HDL implementation via Ethernet. Finally, the control manager in the service provider will send to the development to synthesize, and if no error, to place it in optimal vRRs for bitstream generation. This is the development stage. In the second stage, if the user wants to run his user IP, the generated bitstream will be forwarded to the bitstream library. If not the user gets only the generated bitstream file. The detailes of the flow is given in Figure 6.

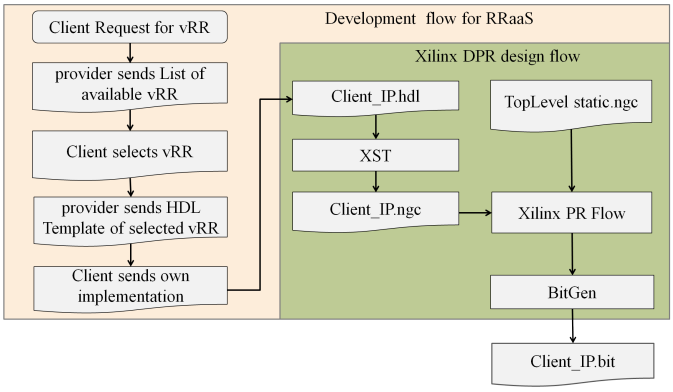


Fig. 6. Flow diagram of RRaaS

To optimize the resource utilization, the size of the virtualized regions are grouped into three as small medium and large. That means, and HDL templete file of the RR region is suitable to the three groups. This will help the control manager to place the user IP in optimal region.

C. Hypervisor

Normally the provision of virtualized machines has three layers; the application layer, hypervisor and the physical layer. The application layer in this case corresponds to the Openstack API-agent interface software. The hypervisor in virtualized FPGA corresponds to the static region or module of the top level structure of the FPGA. Then, the virtualized reconfigurable regions will be instantiated and controlled by the hypervisor. In other words, the hypervisor is the one that manage the virtualized FPGA resources.

V. Experimental results

To test the proposed algorithm, we have implemented image analysis and mathematical manipulation modules and defined three reconfigurable regions for virtualization in the FPGA as shown in Figure 7.. A 2x2-mesh NoC generated from the Atlas NoC generator [24] is used as communication medium. The first router is directely connected to the control/hypervisor and the rest three routers to the virtualized reconfigurable regions. The second router is connected vRR1 where vRR1 is designed for scalar manipulation and includes three modes; addition, subtraction and multiplication. The third router is also connected to vRR2 where this virtulized region is designed for accelerating image processing modes: median filtering, Gray scale and contrasting. Finally, the last router is connected to vRR3 where vRR3 is dedicated for accelerating 8x8 matrix addition and subtraction.

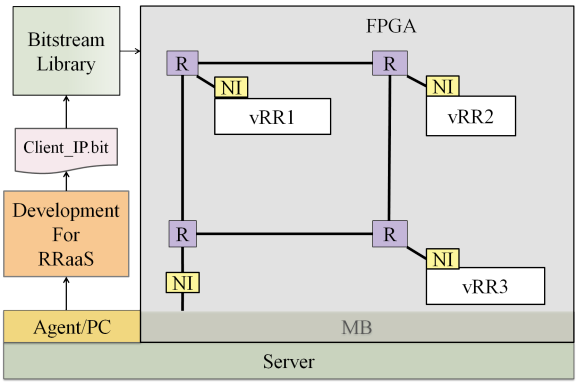


Fig. 7. 2x2 NoC based reconfigurable IPs

A performance evaluation in terms of resource utilizationa and reconfiguration time for the virtualized reconfigurable regions is given in Table-1. The result shows that that the size of the bitstream is too small and generating as much as possible of reconfigurable modules will not cost memory. The reconfiguration time is also to small which is applicable for real time.

TABLE I
PERFORMANCE EVALUATION

DPR Module	Resources Utilization				Bitstream size	config. time
	LUT	FF	DSP	BRAM		
vRR1	2013	1208	12	1	17 KB	0.21 ms
vRR2	9193	7470	40	1	34 KB	0.43 ms
vRR3	5419	6121	40	8	29 KB	0.39 ms

TABLE II
RESOURCE UTILIZATION OF THE 2X2 NOC

LUT	FF	DSP	BRAM
6117	2958	3047	6

As the services provided by a cloud computing must be measurable, there should be continuous status communication between the virtualized reconfigurable regions and the hypervisor which control and manages the resources utilization. It also helps to replace idle accelerators by the blank bitstreams so that power can be saved. It is observed that loading a blank bitstream during idle time saves about 35% of the power

consumption when a real accelerator is loaded and is not performing anything. The NoC based communication helps the structure to have parallel communication and perform both data and control signal communication at the same time.

VI. Conclusions

We have implemented a NoC based virtualized FPGA and tested locally to share the FPGA resources on cloud context. Integration of hardware accelerators in the cloud environment helps to provide FPGA resource in a cheap price and improve the utilization of the FPGA. Similarly, the performance of a cloud environment will be maximized by integrating hardware accelerator FPGA.

The NoC will provide a flixible and scalable parallel communication architecture to the virtualized FPGA resources so that both data and control signal communication can be performed in parallel. We have presented the advantage of integrating a NoC based reconfigurable accelerators in the cloud computing. Two possible virtualized FPGA cloud based service models are proposed in this paper. The first service model, HAaaS, helps the cloud provider to give accelerating services. Whereas the second service model, RRaaS, provides both development and accelerating service. As a future work, the NoC will be extended into dynamically reconfigurable and then it will be deplyed into a server to test it via API.

References

[1] N. Bharathi and P. Neelamegam, "A reconfigurable framework for cloud computing architecture," *Artificial Intelligence*, vol. 6, pp. 117–120, 2013.

[2] A. Huth and J. Cebula, "The Basics of Cloud Computing," *United States Computer Emergency Readiness Team*, pp. 1–4, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780124059320000074

[3] L. Schubert, K. G. Jeffery, and B. NeideckerLutz, "The Future of Cloud Computing: Opportunities for European Cloud Computing Beyond 2010:-expert Group Report," Tech. Rep. European Commission, Information Society and Media, 2010.

[4] T. El-Ghazawi, E. El-Araby, M. Huang, K. Gaj, V. Kindratenko, and D. Buell, "The promise of high-performance reconfigurable computing," *Computer*, vol. 41, no. 2, pp. 69–76, 2008.

[5] Openstack. [Online]. Available: http://www.openstack.org/software/

[6] A. Putnam, A. M. Caulfield, E. S. Chung, D. Chiou, K. Constantinides, J. Demme, H. Esmaeilzadeh, J. Fowers, G. P. Gopal, J. Gray, M. Haselman, S. Hauck, S. Heil, A. Hormati, J.-Y. Kim, S. Lanka, J. Larus, E. Peterson, S. Pope, A. Smith, J. Thong, P. Y. Xiao, and D. Burger, "A reconfigurable fabric for accelerating large-scale datacenter services," in *Proceeding of the 41st Annual International Symposium on Computer Architecuture*, ser. ISCA '14. Piscataway, NJ, USA: IEEE Press, 2014, pp. 13–24. [Online]. Available: http://dl.acm.org/citation.cfm?id=2665671.2665678

[7] S. Byma, J. Steffan, H. Bannazadeh, A. Leon-Garcia, and P. Chow, "Fpgas in the cloud: Booting virtualized hardware accelerators with openstack," in *Field-Programmable Custom Computing Machines (FCCM), 2014 IEEE 22nd Annual International Symposium on*, May 2014, pp. 109–116.

[8] J. Dondo Gazzano, F. Sanchez Molina, F. Rincon, and J. C. López, "Integrating reconfigurable hardware-based grid for high performance computing," *The Scientific World Journal*, vol. 2015, 2015.

[9] Openstack, *OpenStack Operations Guide*. [Online]. Available: http://docs.openstack.org/ops-guide/index.html

[10] R. Dafali, J.-P. Diguet, and M. Sevaux, "Key research issues for reconfigurable network-on-chip," in *Reconfigurable Computing and FPGAs, 2008. ReConFig '08. International Conference on*, Dec 2008, pp. 181–186.

[11] Xilinx, "Partial Reconfiguration User Guide," vol. 702, 2013.

[12] D. Atienza, F. Angiolini, S. Murali, A. Pullini, L. Benini, and G. D. Micheli, "Network-on-chip design and synthesis outlook," *Integration, the {VLSI} Journal*, vol. 41, no. 3, pp. 340 – 359, 2008.

[13] A. Agarwal, C. Iskander, and R. Shankar, "Survey of network on chip (noc) architectures & contributions," *Journal of engineering, Computing and Architecture*, vol. 3, no. 1, pp. 21–27, 2009.

[14] P. Pande, C. Grecu, A. Ivanov, R. Saleh, and G. De Micheli, "Design, synthesis, and test of networks on chips," *Design Test of Computers, IEEE*, vol. 22, no. 5, pp. 404–413, Sept 2005.

[15] É. Cota, A. Morais Amory, and M. Soares Lubaszewski, *Reliability, Availability and Serviceability of Networks-on-Chip*. Springer, 2012.

[16] F. Alshuwaier, A. Alshwaier, and A. Areshey, "Applications of cloud computing in education," in *Computing and Networking Technology (ICCNT), 2012 8th International Conference on*, Aug 2012, pp. 26–33.

[17] P. Francisco, "The Netezza data appliance architecture: A platform for high performance data warehousing and analytics," *IBM Redbooks*, 2011.

[18] A. Putnam, A. Caulfield, E. Chung, D. Chiou, K. Constantinides, J. Demme, H. Esmaeilzadeh, J. Fowers, G. Gopal, J. Gray, M. Haselman, S. Hauck, S. Heil, A. Hormati, J.-Y. Kim, S. Lanka, J. Larus, E. Peterson, S. Pope, A. Smith, J. Thong, P. Xiao, and D. Burger, "A reconfigurable fabric for accelerating large-scale datacenter services," in *Computer Architecture (ISCA), 2014 ACM/IEEE 41st International Symposium on*, June 2014, pp. 13–24.

[19] J. Ouyang, S. Lin, W. Qi, Y. Wang, B. Yu, and S. Jiang, "SDA : Software-Defined Accelerator for Large- Scale DNN Systems," in *HOT CHIPS*, 2014, pp. 1–23.

[20] S. A. Fahmy, K. Vipin, and S. Shreejith, "Virtualized fpga accelerators for efficient cloud computing," in *Cloud Computing Technology and Science (CloudCom), 2015 IEEE 7th International Conference on*, Nov 2015, pp. 430–435.

[21] E. El-Araby, I. Gonzalez, and T. El-Ghazawi, "Virtualizing and sharing reconfigurable resources in high-performance reconfigurable computing systems," in *High-Performance Reconfigurable Computing Technology and Applications, 2008. HPRCTA 2008. Second International Workshop on*, Nov 2008, pp. 1–8.

[22] O. Knodel and R. Spallek, "Computing framework for dynamic integration of reconfigurable resources in a cloud," in *Digital System Design (DSD), 2015 Euromicro Conference on*, Aug 2015, pp. 337–344.

[23] H. L. Kidane, E. Bourennane, and G. Ochoa-Ruiz, "Noc based virtualized accelerators for cloud computing," in *Embedded Multicore/Many-core Systems-on-Chip (MCSoC), 2016 IEEE 10th International Symposium on*, Sept 2016, pp. 133–137.

[24] F. Moraes, N. Calazans, A. Mello, L. Mller, and L. Ost, "Hermes: an infrastructure for low area overhead packet-switching networks on chip," *Integration, the {VLSI} Journal*, vol. 38, no. 1, pp. 69 – 93, 2004.

Collaborative localization and formation flying using distributed stereo-vision

Nathan Piasco, Julien Marzat, Martial Sanfourche

***Abstract*— This paper considers collaborative stereo-vision as a mean of localization for a fleet of micro-air vehicles (MAV) equipped with monocular cameras, inertial measurement units and sonar sensors. A sensor fusion scheme using an extended Kalman filter is designed to estimate the positions and orientations of all the vehicles from these distributed measurements. The estimation is completed by a formation control to maximize the overlapping fields of view of the vehicles. Experimental tests for the complete perception and control loop have been performed on multiple MAVs with centralized processing on a ROS ground station.**

***Index Terms*— collaborative localization, formation control, micro-air vehicles, stereo-vision**

I. Introduction

Vision-based localization is a popular approach in the field of robotics, particularly when the robots considered are flying vehicles in GPS-denied cluttered environments. Numerous SLAM and visual odometry methods have been developed to localize a single aerial vehicle, based on either monocular data [1], [2], which suffers from depth uncertainty, or stereo-vision [3], [4], which uses additional 3D information.

Recent research has focused on fleets of MAVs, whose main interests are to carry complementary sensors on cheaper vehicles as well as their ability to cover more field [5]. The context considered is to localize all MAVs in a fleet by fusing their distributed embedded measurements. A first approach of collaborative localization consists in merging individual maps created by mono-vehicle monocular SLAMs. The algorithms from [6]–[8] propose efficient fusion strategies, however this approach suffers from the inherent drawbacks of monocular SLAMs, i.e. depth uncertainty and drift. As an alternative approach, the fusion of collaborative stereo-vision (with a varying baseline, as each vehicle moves) with IMU data has been investigated in [9] for estimating the relative pose of two MAVs. In [10], the relative localization of multiple MAVs was obtained by combining IMU measurements and an homography estimation. The same authors also proposed a formation control to maintain the fleet in a desired layout. These papers contain promising results on stereo-vision for fleet localization, however the entire estimation process was not tested in the experiments reported since the vision algorithms were emulated by motion tracking data.

The present work proposes a filtering scheme (Section II) to localize in a global frame all the MAVs of a fleet, using their monocular cameras in a collaborative stereo-vision process and IMUs but also additional altitude measurements and linear velocity estimates. A formation control law is also proposed (Section III) for maximizing the overlap of the MAV fields of view, in order to enhance cooperative localization. The complete vision and control loop has been flight-tested on multiple Parrot® AR Drones [11] with centralized processing on a ROS ground station.

N. Piasco is a Research Engineer at A.I.Mergence, F-75013 Paris, France nathan.piasco@ai-mergence.com

J. Marzat and M. Sanfourche are Research Scientists at ONERA – The French Aerospace Lab, F-91123 Palaiseau, France, julien.marzat@onera.fr, martial.sanfourche@onera.fr

II. Collaborative localization

A. Notations

A vector ${}^{r}\boldsymbol{x}_i$ stands for a variable associated with vehicle i and expressed in the coordinate frame linked to vehicle r. The world inertial frame is denoted by w. Rotations are represented by quaternions according to *Hamilton* notation:

$$\overline{q} = \begin{bmatrix} q \\ \boldsymbol{q} \end{bmatrix} \tag{1}$$

where q stands for the real part of the quaternion and the vector $\boldsymbol{q}$ is associated with the imaginary part. *Hamilton* product of two quaternions is defined as follows [12]:

$$\overline{q}_1 \otimes \overline{q}_2 = \begin{bmatrix} q_1 q_2 - \boldsymbol{q_1} \cdot \boldsymbol{q_2} \\ q_1 \boldsymbol{q_2} + q_2 \boldsymbol{q_1} + \boldsymbol{q_1} \times \boldsymbol{q_2} \end{bmatrix} \tag{2}$$

where $\cdot$ refers to the dot product and $\times$ the cross product. Measurements and variables symbols used afterward are:

$\boldsymbol{p}$ for a position,
$\boldsymbol{\omega}$ for an angular velocity,
$\boldsymbol{v}$ for a linear velocity,
$\boldsymbol{R}(\overline{q})$ for a rotation matrix associated with $\overline{q}$.

B. MAV characteristics and sensors

Multiple Parrot® AR Drones 2.0 have been used to test the filter and the control law described in this paper. These low-cost vehicles are equipped with:

- two monocular cameras: one at the front of the MAV and the other downward,
- an IMU (3-axis accelerometers and 3-axis gyrometers),
- a sonar sensor directed downward,
- an on-board non-programmable CPU.

The communication between the MAVs and a laptop is established through a Wi-Fi connection. The MAV driver provides a pre-filtered information about linear velocity of the MAV, computed by sensor fusion between optical flow acquired by the bottom camera and IMU accelerations [11]. The sonar sensor provides an altitude measurement in the MAV frame, ${}^{i}z_i$. The MAV altitude in world coordinates can be obtained by projecting this value as ${}^{w}z_i = \boldsymbol{R}({}^{w}\overline{q}_i){}^{i}z_i$.

The following measurements are thus available as inputs for each vehicle in the filtering scheme:

- angular velocity ${}^i\boldsymbol{\omega}_i$,
- linear velocity ${}^i\boldsymbol{v}_i$,
- ground altitude ${}^w z_i$,
- raw front image (640x360 resolution).

The state vector (pose) of the i-th MAV is

$$\boldsymbol{x}_i = \begin{pmatrix} {}^w\overline{q}_i \\ {}^w\boldsymbol{p}_i \end{pmatrix} \tag{3}$$

and its dynamics are modeled by

$$\begin{array}{l} {}^w\dot{\overline{q}}_i = \frac{1}{2}\left({}^w\overline{q}_i \otimes {}^i\overline{\omega}_i\right) \\ {}^w\dot{\boldsymbol{p}}_i = \boldsymbol{R}({}^w\overline{q}_i)\boldsymbol{v}_i \end{array}, \tag{4}$$

where ${}^i\overline{\omega}_i = \begin{pmatrix} 0 & {}^i\boldsymbol{\omega}_i^T \end{pmatrix}^T$. This model assumes that faster loops regulate the orientation and velocity such that direct velocity control inputs can be applied (which is allowed by the ardrone_autonomy ROS driver[1]).

C. Distributed stereo-vision

The aim of visual pose recognition is to compute the relative pose between two cameras with known intrinsic parameters and overlapping fields of view. The estimated pose contains:

- the relative rotation from camera 2 to camera 1

$${}^1\overline{q}_2{}^v \equiv {}^1\overline{q}_2, \tag{5}$$

- the relative position, up to a scale factor λ, of camera 2 to camera 1: ${}^1\boldsymbol{p}_2{}^v \equiv \lambda {}^1\boldsymbol{p}_2$. To deal with this scale factor in the filtering scheme, the estimated translation is normalized such that

$${}^1\boldsymbol{p}_2{}^v = {}^1\boldsymbol{p}_2 \| {}^1\boldsymbol{p}_2 \|^{-1}. \tag{6}$$

The usual approach involves to first establish feature correspondences between the two images (features f_i from image 1 and f_i' from image 2 correspond to 3D points p_i) and then solve the pose recognition problem.

1) Matching method: The matching process can be divided in two steps: an extraction of features and description on each image, followed by descriptor pairing. The SIFT [13] feature points extractor and descriptor was used (OpenCV implementation). This widely-used algorithm is versatile and scale-invariant, which is important in the case of large baseline stereo-vision problems.

2) Relative pose computation: Once the feature correspondences have been established, the five-point algorithm [14] is used to recover the relative pose between the two cameras (OpenGV [15] implementation). Nistér's algorithm computes the essential matrix $\boldsymbol{E}$ that contains the relative rotation and translation information $\boldsymbol{E} = \boldsymbol{R}[\boldsymbol{t}]_\times$, where $[\boldsymbol{t}]_\times$ denotes a skew-symmetric matrix of $\boldsymbol{t}$,

$$[\boldsymbol{t}]_\times = \begin{pmatrix} 0 & -t_x & t_y \\ t_z & 0 & -t_z \\ -t_y & t_x & 0 \end{pmatrix}. \tag{7}$$

[1]https://github.com/AutonomyLab/ardrone_autonomy

An *SVD* decomposition of $\boldsymbol{E}$ makes it possible to recover the relative rotation matrix $\boldsymbol{R}$ and translation vector $\boldsymbol{t}$. As the matching part provides most of the time more than five corresponding features, a RANSAC step was introduced to minimize the re-projection error of the feature points on the essential matrix [15]. In the experiments, the computation time of the pose reconstruction process (synchronized images) was always lower than 0.5s on a laptop with an Intel® i7 processor, while the filter frequency was set to 15Hz.

D. Filtering scheme

This section describes the sensor fusion scheme, which consists in an Extended Kalman Filter (EKF), to estimate positions and orientations of all the MAVs.

1) Two-vehicle case: Before considering a complete fleet of MAVs, the filter design is presented for estimating poses of two vehicles in the world frame. The state vector considered contains the poses of the two MAVS as

$$\boldsymbol{X} = \begin{pmatrix} {}^w\overline{q}_1^T & {}^w\overline{q}_2^T & {}^w\boldsymbol{p}_1^T & {}^w\boldsymbol{p}_2^T \end{pmatrix}^T. \tag{8}$$

Unlike [9], the scale factor λ resulting from the stereo-vision process is not needed here, due to the normalization from equation (6). The input vector is composed of the variables collected from the IMUs and embedded velocity filters:

$$\boldsymbol{U} = \begin{pmatrix} {}^1\overline{\omega}_1^T & {}^2\overline{\omega}_2^T & {}^1\boldsymbol{v}_1^T & {}^2\boldsymbol{v}_2^T \end{pmatrix}^T. \tag{9}$$

The measurement vector incorporates the relative pose computed by the stereo-vision process and the absolute ground altitudes measured by the sonar sensors:

$$\boldsymbol{Y} = \left(({}^1\boldsymbol{p}_2{}^v)^T \ \ ({}^1\overline{q}_2{}^v)^T \ \ {}^w z_1 \ \ {}^w z_2 \right)^T. \tag{10}$$

It is assumed that the distance from the camera to the IMU is small enough to be negligible. Deriving the state vector leads to the following formulation:

$$\begin{array}{rcl} \dot{\boldsymbol{X}} & = & f(\boldsymbol{X}, \boldsymbol{U} + \boldsymbol{W}) \\ \dot{\boldsymbol{X}} & = & \begin{pmatrix} {}^w\dot{\overline{q}}_1 \\ {}^w\dot{\overline{q}}_2 \\ {}^w\dot{\boldsymbol{p}}_1 \\ {}^w\dot{\boldsymbol{p}}_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left({}^w\overline{q}_1 \otimes ({}^1\overline{\omega}_1 + \boldsymbol{\eta}_{\overline{\boldsymbol{\omega}}})\right) \\ \frac{1}{2}\left({}^w\overline{q}_2 \otimes ({}^2\overline{\omega}_2 + \boldsymbol{\eta}_{\overline{\boldsymbol{\omega}}})\right) \\ \boldsymbol{R}({}^w\overline{q}_1)({}^1\boldsymbol{v}_1 + \boldsymbol{\eta}_{\boldsymbol{v}}) \\ \boldsymbol{R}({}^w\overline{q}_2)({}^2\boldsymbol{v}_2 + \boldsymbol{\eta}_{\boldsymbol{v}}) \end{pmatrix} \end{array} \tag{11}$$

where $\boldsymbol{W} = (\boldsymbol{\eta}_{\overline{\boldsymbol{\omega}}}^T \boldsymbol{\eta}_{\overline{\boldsymbol{\omega}}}^T \boldsymbol{\eta}_{\boldsymbol{v}}^T \boldsymbol{\eta}_{\boldsymbol{v}}^T)^T$ is the zero-mean Gaussian input noise vector, whose variances are reported on the diagonal of the input covariance matrix $\boldsymbol{Q}$. The measurement vector can be expressed according to the state variables as

$$\begin{array}{rcl} \boldsymbol{Y} & = & h(\boldsymbol{X}) + \boldsymbol{V} \\ & = & \begin{pmatrix} \boldsymbol{R}({}^w\overline{q}_1)\left({}^w\boldsymbol{p}_1 - {}^w\boldsymbol{p}_2\right)\lambda_{12} \\ {}^w\overline{q}_1^* \otimes {}^w\overline{q}_2 \\ {}^w p_{1,z} \\ {}^w p_{2,z} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_{\boldsymbol{p_r}} \\ \boldsymbol{\eta}_{\boldsymbol{q_r}} \\ \eta_z \\ \eta_z \end{pmatrix} \end{array} \tag{12}$$

where:

- $\boldsymbol{V} = \left(\boldsymbol{\eta}_{\boldsymbol{p_r}}^T \boldsymbol{\eta}_{\boldsymbol{q_r}}^T \eta_z \, \eta_z\right)^T$ is the zero-mean Gaussian measurement noise vector, whose variances are placed on the diagonal of the output covariance matrix $\boldsymbol{R}$,

- ${}^w p_{i,z}$ stands for the third component of vector ${}^w\boldsymbol{p}_i$,
- ${}^w\overline{q}_1^*$ represents conjugate quaternion of ${}^w\overline{q}_1$ [9],
- the scale factor is taken equal to the inverse of the distance between the MAVs, $\lambda_{ij} = \|{}^w\boldsymbol{p}_i - {}^w\boldsymbol{p}_j\|^{-1}$. The predicted positions are used and considered as constant in the filter derivation. This approximation performed better in preliminary tests than introducing the exact linearization, which showed numerical stability issues.

The Euler discretization of equation (11) yields

$$\begin{aligned} \boldsymbol{X}_{k+1/k} &= \boldsymbol{X}_{k/k} + f(\boldsymbol{X}_{k/k}, \boldsymbol{U}_k)\Delta t \\ &= \tilde{f}(\boldsymbol{X}_{k/k}, \boldsymbol{U}_k). \end{aligned} \tag{13}$$

Kalman equations give the evolution of the matrix covariance error attached to the state vector,

$$\boldsymbol{P}_{k+1/k} = \boldsymbol{F}_k \boldsymbol{P}_{k/k} \boldsymbol{F}_k^T + \boldsymbol{L}_k \boldsymbol{Q} \boldsymbol{L}_k^T \tag{14}$$

$$\text{where } \boldsymbol{F}_k = \left.\frac{\partial \tilde{f}}{\partial \boldsymbol{X}}\right|_{\boldsymbol{X}_{k/k}, \boldsymbol{U}_k}, \quad \boldsymbol{L}_k = \left.\frac{\partial \tilde{f}}{\partial \boldsymbol{U}}\right|_{\boldsymbol{X}_{k/k}, \boldsymbol{U}_k}. \tag{15}$$

The state vector is updated with the $\boldsymbol{Y}_k$ vector:

$$\begin{aligned} \boldsymbol{K}_f &= \boldsymbol{P}_{k+1/k} \boldsymbol{H}_k^T (\boldsymbol{H}_k \boldsymbol{P}_{k+1/k} \boldsymbol{H}_k^T + \boldsymbol{R})^{-1} \\ \boldsymbol{X}_{k+1/k+1} &= \boldsymbol{X}_{k+1/k} + \boldsymbol{K}_f \left(\boldsymbol{Y}_k - h(\boldsymbol{X}_{k+1/k})\right) \\ \boldsymbol{P}_{k+1/k+1} &= (\boldsymbol{I} - \boldsymbol{K}_f \boldsymbol{H}_k) \boldsymbol{P}_{k+1/k} \end{aligned} \tag{16}$$

where $\boldsymbol{K}_f$ is the Kalman gain and $\boldsymbol{H}_k = \left.\frac{\partial h}{\partial \boldsymbol{X}}\right|_{\boldsymbol{X}_{k/k}}$.

2) General case: The EKF can be generalized for a fleet of N MAVs. The state vector becomes an $7N$-long vector,

$$\boldsymbol{X} = \left({}^w\overline{q}_1^T \ \ldots \ {}^w\overline{q}_N^T \ {}^w\boldsymbol{p}_1^T \ \ldots \ {}^w\boldsymbol{p}_N^T\right)^T \tag{17}$$

and the input vector can now be written as

$$\boldsymbol{U} = \left({}^1\overline{\omega}_1^T \ \ldots \ {}^N\overline{\omega}_N^T \ {}^1\boldsymbol{v}_1^T \ \ldots \ {}^N\boldsymbol{v}_N^T\right)^T. \tag{18}$$

The measurement vector now depends on the two MAVs i and j whose images have been used for computing relative pose information, as well as all altitude measurements:

$$\boldsymbol{Y}_{i,j} = \left(({}^i\boldsymbol{p}_j{}^v)^T \ ({}^i\overline{q}_j{}^v)^T \ {}^w z_1 \ \ldots \ {}^w z_N\right)^T \tag{19}$$

The state and output equations are now

$$\dot{\boldsymbol{X}} = f(\boldsymbol{X}, \boldsymbol{U} + \boldsymbol{W}) = \begin{pmatrix} \frac{1}{2}\left({}^w\overline{q}_1 \otimes ({}^1\overline{\omega}_1 + \boldsymbol{\eta}_{\overline{\omega}})\right) \\ \vdots \\ \frac{1}{2}\left({}^w\overline{q}_N \otimes ({}^N\overline{\omega}_N + \boldsymbol{\eta}_{\overline{\omega}})\right) \\ \boldsymbol{R}({}^w\overline{q}_1)({}^1\boldsymbol{v}_1 + \boldsymbol{\eta}_{\boldsymbol{v}}) \\ \vdots \\ \boldsymbol{R}({}^w\overline{q}_N)({}^N\boldsymbol{v}_N + \boldsymbol{\eta}_{\boldsymbol{v}}) \end{pmatrix} \tag{20}$$

$$\boldsymbol{Y}_{i,j} = h_{i,j}(\boldsymbol{X}) + \boldsymbol{V} = \begin{pmatrix} \boldsymbol{R}({}^w\overline{q}_i)({}^w\boldsymbol{p}_i - {}^w\boldsymbol{p}_j)\lambda_{ij} + \boldsymbol{\eta}_{\boldsymbol{p_r}} \\ {}^w\overline{q}_i^* \otimes {}^w\overline{q}_j + \boldsymbol{\eta}_{\boldsymbol{q_r}} \\ {}^w p_{1,z} + \eta_z \\ \vdots \\ {}^w p_{N,z} + \eta_z \end{pmatrix} \tag{21}$$

Each update step of the EKF corrects solely the poses of two different MAVs:

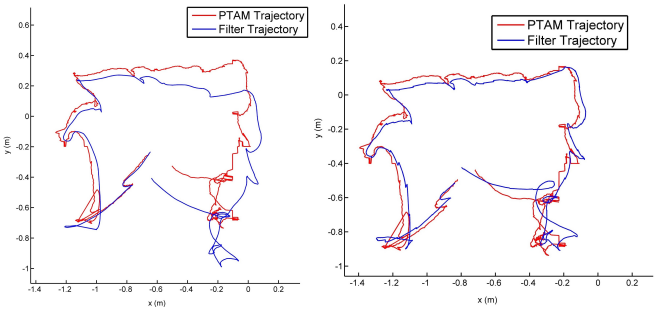
$$\begin{aligned} \boldsymbol{K}_f &= \boldsymbol{P}_{k+1/k} \boldsymbol{H}_{i,j,k}^T \left(\boldsymbol{H}_{i,j,k} \boldsymbol{P}_{k+1/k} \boldsymbol{H}_{i,j,k}^T + \boldsymbol{R}\right)^{-1} \\ \boldsymbol{X}_{k+1/k+1} &= \boldsymbol{X}_{k+1/k} + \boldsymbol{K}_f \left(\boldsymbol{Y}_{i,j} - h_{i,j}(\boldsymbol{X}_{k+1/k})\right) \\ \boldsymbol{P}_{k+1/k+1} &= (\boldsymbol{I} - \boldsymbol{K}_f \boldsymbol{H}_{i,j,k}) \boldsymbol{P}_{k+1/k} \end{aligned} \tag{22}$$

with $\boldsymbol{H}_{i,j,k} = \left.\frac{\partial h_{i,j}}{\partial \boldsymbol{X}}\right|_{\boldsymbol{X}_{k/k}}$.

If the pairs of MAVs used in the stereo-vision process are chosen appropriately at each time step (see Section III-A), the poses of all vehicles of the fleet can be updated.

E. Experimental results

The proposed filter has been compared with the monocular AR Drone SLAM described in [2], which is related to the parallel tracking and mapping method (PTAM) [1]. During this experiment with two MAVs, one of the vehicles followed a reference trajectory to draw a one-meter-side square while the other was hovering. The localization computed by our filter, with and without stereo-vision correction, was compared with this monocular method for the moving vehicle (Figure 1, Table I).The results show that the stereo-vision correction significantly improves the localization, with a smaller mean-square error, and the obtained trajectory is consistent with the monocular SLAM (which is however unable to deal with multiple vehicles). Successful experiments have also been conducted for a fleet of three MAVs.



(a) Only-predictive filter (b) Filter with stereo-vision correction

Fig. 1: Comparison of the filter with monocular SLAM [2]

TABLE I: Filter evaluation (MSE)

Method	x-axis	y-axis
Only-predictive filter	$0.01294\,\mathrm{m}^2$	$0.01522\,\mathrm{m}^2$
Filter with stereo-vision correction	$0.00477\,\mathrm{m}^2$	$0.00456\,\mathrm{m}^2$

An experimental observability analysis has been conducted by studying the rank of the following $(7+N)(7N) \times (7N)$ observability matrix during the estimation process:

$$\boldsymbol{Obs}_k = \begin{pmatrix} \boldsymbol{H}_k \\ \boldsymbol{H}_k \boldsymbol{F}_{k-7N-1} \\ \boldsymbol{H}_k \boldsymbol{F}_{k-7N-1} \boldsymbol{F}_{k-7N-2} \\ \vdots \\ \boldsymbol{H}_k \boldsymbol{F}_{k-7N-1} \boldsymbol{F}_{k-7N-2} \cdots \boldsymbol{F}_k \end{pmatrix} \tag{23}$$

The results showed that the rank of $\boldsymbol{Obs}_k$ is always 2 degrees below its number of columns. Indeed, the lateral translations are not fully observable due to relative pose correction (unlike the vertical axis thanks to the altitude sensor). The filter thus behaves as an odometry on these axes, which requires a good initialization to remain accurate.

III. Formation control

In this section, a formation control is proposed to simultaneously (i) maximize the overlapping fields of view of MAV pairs to guarantee stereo-vision feasibility and (ii) ensure collision avoidance between MAVs. Each MAV has four control inputs: three linear velocities and a yaw angular velocity. To simplify the problem, the control law is divided into three independent parts (see Figure 2):

1) align the MAVs on the same reference yaw,
2) constrain the vehicles in a common reference vertical plane $\mathcal{P}$,
3) maximize the area of overlapping fields of view by pairs of MAVs and avoid collisions.

The first two parts are achieved by simple proportional controllers (for fleet motion, the yaw and vertical plane references can be modified or associated with a MAV leader). The third objective requires a more complex control law, detailed in what follows, to compute the linear velocities $\boldsymbol{u}_i = (u_y\, u_z)^T$ where the y and z axes form an orthogonal basis in the reference plane $\mathcal{P}$.

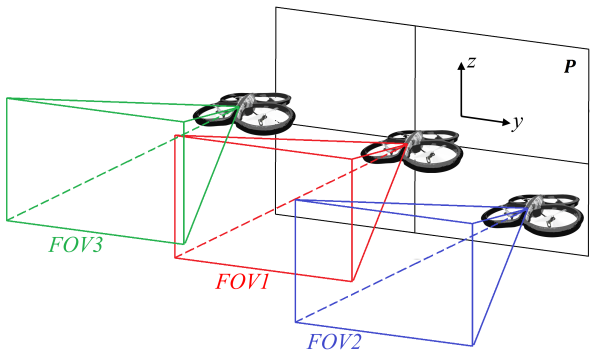


Fig. 2: MAVs are headed in the same direction, in a common vertical plane. The control law aims at maximizing pairs of overlapping fields of view FOV_1, FOV_2 and FOV_3.

A. Maximizing overlapping fields of view

A camera field of view (FOV) can be described as shown in Figure 3, with its width and height parameterized by the depth of field d and the camera angles of view α and β,

$$\begin{cases} w = 2d\tan\left(\dfrac{\alpha}{2}\right) \\ h = 2d\tan\left(\dfrac{\beta}{2}\right) \end{cases}. \tag{24}$$

With this parameterization, the area of overlapping FOVs for MAV pairs can be computed as a function of the MAV positions in the vertical plane, and maximized locally thanks to a gradient ascent algorithm. An algorithm is proposed to compute this area A_Σ, which is basically the sum of the areas of hatched rectangles in Figure 4.

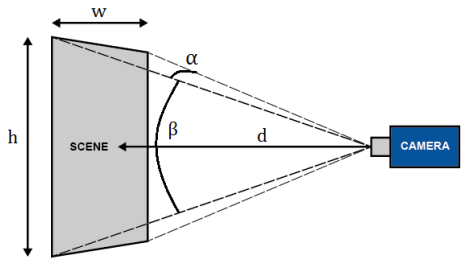


Fig. 3: Field of view of a camera

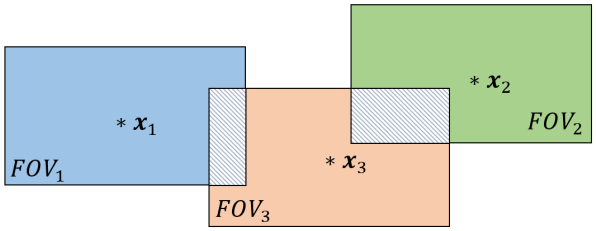


Fig. 4: Area of overlapping fields of view, to be maximized

The area $A_{i,j}$ of the overlap of two MAV FOVs i and j is equal to

$$\begin{aligned} A_{i,j} = &(\min(x_{i,y}, x_{j,y}) - \max(x_{i,y}, x_{j,y}) + w) \\ &\cdot (\min(x_{i,z}, x_{j,z}) - \max(x_{i,z}, x_{j,z}) + h) \end{aligned} \tag{25}$$

where $\boldsymbol{x}_i$ is the position of MAV i in $\mathcal{P}$,

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i,y} \\ x_{i,z} \end{pmatrix}. \tag{26}$$

This quantity exists only if the following conditions are true:

$$\begin{cases} \|x_{i,y} - x_{j,y}\| < w \\ \|x_{i,z} - x_{j,z}\| < h \end{cases}, \tag{27}$$

otherwise $A_{i,j}$ is set to 0. The area function A_Σ can be computed using Algorithm 1, which chooses the minimum number of MAV pairs to have a maximum area of FOV overlaps, while ensuring that all vehicles are taken into account (see Figure 5). This also provides the minimal number of image pairs on which the stereo-vision should be computed, while preserving the connectivity of the fleet.

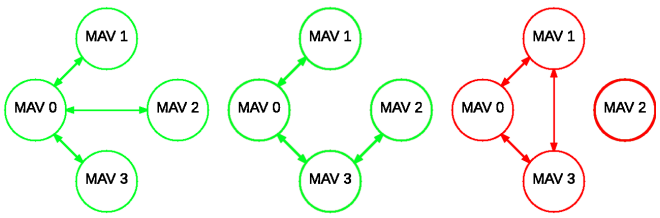


Fig. 5: Algorithm 1 output: last configuration is forbidden

Algorithm 1: Computation of overlapping area function $A_\Sigma(\boldsymbol{x})$

Input : all vehicle positions $\boldsymbol{x}_i, i \in [1, N]$
Output: A_Σ
Data: L initially empty list
Areas for all possible pairs with overlapping FOVs
for $i \leftarrow 1$ **to** $N-1$ **do**
  for $j \leftarrow i+1$ **to** N **do**
    if $\|x_{i,y} - x_{j,y}\| < w$ *and* $\|x_{i,z} - x_{j,z}\| < h$ **then**
      add $A_{i,j}$ to L;
    end
  end
end
$A_\Sigma \leftarrow 0$;
A_Σ is the sum of the $N-1$ largest $A_{i,j}$ such that each MAV is included
for $i \leftarrow 1$ **to** $N-1$ **do**
  $A_\Sigma \leftarrow A_\Sigma + \max L$;
  remove $\max L$ from L;
end

B. Collision avoidance

The following repulsive term, described in [16], was used to prevent inter-vehicles collisions:

$$\boldsymbol{u}_i^{\text{col}} = 2k \sum_{j=1}^{N} (\boldsymbol{x}_i - \boldsymbol{x}_j) \frac{g_{ij}}{q}, \tag{28}$$

where k is a positive control gain, $g_{ij} = \exp\left(-\boldsymbol{\delta}_{ij}^T \boldsymbol{\delta}_{ij}/q\right)$, with $\boldsymbol{\delta}_{ij} = (\boldsymbol{x}_i - \boldsymbol{x}_j)$ and q a positive constant determining the repulsion distance.

C. Complete control law

The resulting control input in the vertical plane is obtained by summing the attractive term represented by the maximization of the overlapping FOVs and the repulsive term:

$$\boldsymbol{u}_i = 2k \sum_{j=1}^{N} (\boldsymbol{x}_i - \boldsymbol{x}_j) \frac{g_{ij}}{q} + k' \left(\frac{\partial A_\Sigma(\boldsymbol{x})}{\partial \boldsymbol{x}_i} \right)^T \tag{29}$$

where

k' is a positive constant,
$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1^T & \boldsymbol{x}_2^T & \dots & \boldsymbol{x}_N^T \end{pmatrix}^T$,
$\dfrac{\partial A_\Sigma(\boldsymbol{x})}{\partial \boldsymbol{x}_i} = \begin{pmatrix} \dfrac{\partial A_\Sigma(\boldsymbol{x})}{\partial x_{i,y}} & \dfrac{\partial A_\Sigma(\boldsymbol{x})}{\partial x_{i,z}} \end{pmatrix}$.

To analyze the stability of this control law, consider the following Lyapunov function:

$$V = \frac{1}{2} \sum_{i=1}^{N} k \sum_{j=1}^{N} g_{ij} + k'(K - A_\Sigma(\boldsymbol{x})), \tag{30}$$

with K a positive constant such as $K > \max_{\boldsymbol{x}}(A_\Sigma(\boldsymbol{x}))$ to ensure that V is positive. $\dot{V}$ can be expressed as

$$\dot{V} = -k \sum_{i=1}^{N} \sum_{j=1}^{N} \dot{\boldsymbol{\delta}}_{ij}^T \boldsymbol{\delta}_{ij} \frac{g_{ij}}{q} - k' \nabla A_\Sigma(\boldsymbol{x}) \dot{\boldsymbol{x}} \tag{31}$$

with $\nabla A_\Sigma(\boldsymbol{x}) = \begin{pmatrix} \dfrac{\partial A_\Sigma(\boldsymbol{x})}{\partial \boldsymbol{x}_1} & \dfrac{\partial A_\Sigma(\boldsymbol{x})}{\partial \boldsymbol{x}_2} & \dots & \dfrac{\partial A_\Sigma(\boldsymbol{x})}{\partial \boldsymbol{x}_N} \end{pmatrix}$.
Following [17], the double sum can be rewritten as

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \dot{\boldsymbol{\delta}}_{ij}^T \boldsymbol{\delta}_{ij} \frac{g_{ij}}{q} = 2 \sum_{i=1}^{N} \sum_{j=1}^{N} (\dot{\boldsymbol{x}}_i)^T (\boldsymbol{x}_i - \boldsymbol{x}_j) \frac{g_{ij}}{q} \tag{32}$$

Replacing in equation (31) yields

$$\begin{aligned}
\dot{V} &= -2k \sum_{i=1}^{N} \sum_{j=1}^{N} (\dot{\boldsymbol{x}}_i)^T (\boldsymbol{x}_i - \boldsymbol{x}_j) \frac{g_{ij}}{q} - k' \nabla A_\Sigma(\boldsymbol{x}) \dot{\boldsymbol{x}} \\
\dot{V} &= -2k \sum_{i=1}^{N} (\dot{\boldsymbol{x}}_i)^T \sum_{j=1}^{N} (\boldsymbol{x}_i - \boldsymbol{x}_j) \frac{g_{ij}}{q} \\
&\qquad -k' \sum_{i=1}^{N} (\dot{\boldsymbol{x}}_i)^T \left(\frac{\partial A_\Sigma(\boldsymbol{x})}{\partial \boldsymbol{x}_i} \right)^T \\
\dot{V} &= \sum_{i=1}^{N} (\dot{\boldsymbol{x}}_i)^T \left(-2k \sum_{j=1}^{N} (\boldsymbol{x}_i - \boldsymbol{x}_j) \frac{g_{ij}}{q} \right. \\
&\qquad \left. -k' \left(\frac{\partial A_\Sigma(\boldsymbol{x})}{\partial \boldsymbol{x}_i} \right)^T \right)
\end{aligned} \tag{33}$$

Since $\dot{\boldsymbol{x}}_i = \boldsymbol{u}_i$, it follows with equation (29) that

$$\begin{aligned}
\dot{V} &= \sum_{i=1}^{N} \boldsymbol{u}_i^T \left(-2k \sum_{j=1}^{N} (\boldsymbol{x}_i - \boldsymbol{x}_j) \frac{g_{ij}}{q} - k' \left(\frac{\partial A_\Sigma(\boldsymbol{x})}{\partial \boldsymbol{x}_i} \right)^T \right) \\
\dot{V} &= \sum_{i=1}^{N} \left(2k \sum_{j=1}^{N} (\boldsymbol{x}_i - \boldsymbol{x}_j) \frac{g_{ij}}{q} + k' \left(\frac{\partial A_\Sigma(\boldsymbol{x})}{\partial \boldsymbol{x}_i} \right)^T \right)^T \\
&\qquad \cdot \left(-2k \sum_{j=1}^{N} (\boldsymbol{x}_i - \boldsymbol{x}_j) \frac{g_{ij}}{q} - k' \left(\frac{\partial A_\Sigma(\boldsymbol{x})}{\partial \boldsymbol{x}_i} \right)^T \right) \\
\dot{V} &= -\sum_{i=1}^{N} \boldsymbol{u}_i^T \boldsymbol{u}_i \leq 0
\end{aligned}$$

The derivative $\dot{V}$ of the Lyapunov function V is thus negative semi-definite for the designed control law, which guarantees that the MAVs converge locally to an equilibrium. This corresponds to the situation where $\dot{V} = 0$, i.e. the attraction toward the local maximum of A_Σ is exactly compensated by the repulsion force (28) resulting in a null velocity $\boldsymbol{u}_i = 0$.

D. Results

1) Simulation: Simulations have been performed to analyze the behavior of a larger fleet (see Figure 6). As the maximization is gradient-based, the final layout of the fleet depends on the initial positions of the MAVs.

2) Experiment: An experiment has been made with two MAVs running the entire localization and control loop. Results of this test are presented in Figure 7 and 8. It can be seen that the vehicles converge to a stable formation that maximizes the overlap of their FOVs and respects the collision avoidance distance, which was set to 3 m via parameter q.

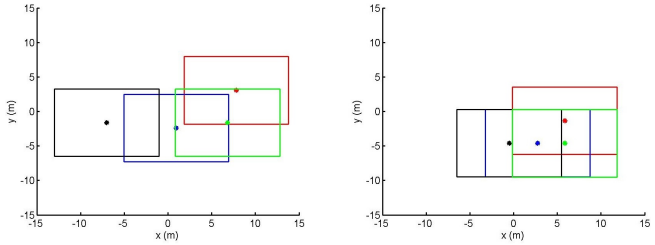


Fig. 6: Simulation with 4 MAVs: positions (dots) and associated FOVs (rectangles) at $t = 0s$ (left) and $t = 10s$ (right).

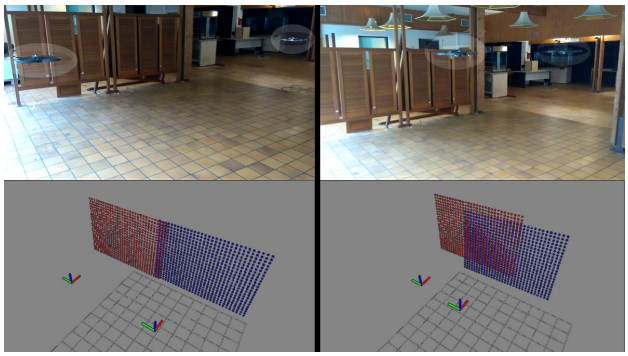


Fig. 7: View of the AR Drones (top) and of the estimated poses (bottom) at $t = 0$s (left) and $t = 8$s (right). Red and blue planes represent the MAV FOVs.

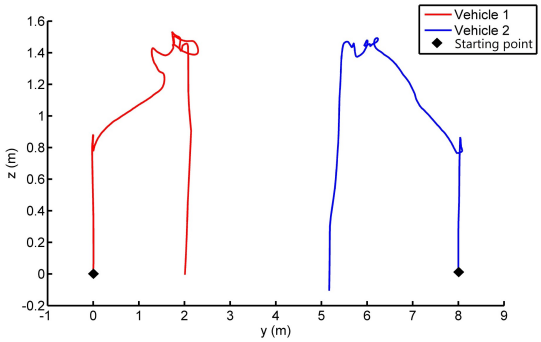


Fig. 8: Experimental MAV trajectories in the y-z plan

IV. Conclusions and Perspectives

In this paper, a sensor fusion scheme including collaborative stereo-vision, IMU, sonar and linear velocity data within an EKF has been proposed to estimate the poses of all the vehicles of a fleet. This algorithm is associated with a control law that maximizes the overlapping fields of view of MAVs in order to improve the stereo-vision process. Experimental results with Parrot® AR Drones show the interest of the complete vision-based estimation and control loop. Stereo-vision was run in real time, which is promising regarding the applicability of this technique for localizing fleets of MAVs in GPS-denied environments.

The localization and control algorithms are designed to incorporate much more than two or three MAVs, which remains to be confirmed by future experimental results. With more MAVs, it could also be interesting to exploit the redundancy of visual sensors to extend the stereo-vision to trifocal (or more) relative pose estimation [18]. Finally, the control law could be extended to perform more elaborate fleet tasks, and the computation load could be distributed if more advanced MAVs are used.

References

[1] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, Nara, Japan*, 2007, pp. 225–234.

[2] J. Engel, J. Sturm, and D. Cremers, "Scale-aware navigation of a low-cost quadrocopter with a monocular camera," *Robotics and Autonomous Systems*, vol. 62, no. 11, pp. 1646–1656, 2014.

[3] J.-H. Sun, B.-S. Jeon, J.-W. Lim, and M.-T. Lim, "Stereo vision based 3D modeling system for mobile robot," in *Proceedings of the International Conference on Control Automation and Systems, Gyeonggi-do, South Korea*, 2010, pp. 71–75.

[4] M. Sanfourche, V. Vittori, and G. L. Besnerais, "eVO: A realtime embedded stereo odometry for MAV applications," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan*, 2013, pp. 2107–2114.

[5] R. M. Murray, "Recent research in cooperative control of multivehicle systems," *Journal of Dynamic Systems, Measurement, and Control*, vol. 129, no. 5, pp. 571–583, 2007.

[6] C. Forster, S. Lynen, L. Kneip, and D. Scaramuzza, "Collaborative monocular SLAM with multiple micro aerial vehicles," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan*, 2013, pp. 3962–3970.

[7] D. Zou and P. Tan, "CoSLAM: Collaborative visual SLAM in dynamic environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 354–366, 2013.

[8] G. Bresson, R. Aufrere, and R. Chapuis, "Consistent multi-robot decentralized SLAM with unknown initial positions," in *Proceedings of the 16th International Conference on Information Fusion, Istanbul, Turkey*, 2013, pp. 372–379.

[9] M. W. Achtelik, S. Weiss, M. Chli, F. Dellaert, and R. Siegwart, "Collaborative stereo," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, USA*, 2011, pp. 2242–2248.

[10] E. Montijano, D. Zhou, E. Cristofalo, M. Schwager, and C. Sagues, "Vision-based distributed formation control without a global reference frame," *International Journal of Robotics Research*, 2014.

[11] P.-J. Bristeau, F. Callou, D. Vissiere, and N. Petit, "The navigation and control technology inside the AR drone micro UAV," in *Proceedings of the 18th IFAC world congress, Milan, Italy*, vol. 18, no. 1, 2011, pp. 1477–1484.

[12] R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[13] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE International Conference on Computer Vision, Kerkyra, Greece*, vol. 2, 1999, pp. 1150–1157.

[14] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 756–770, 2004.

[15] L. Kneip and P. Furgale, "OpenGV: A unified and generalized approach to real-time calibrated geometric vision," in *Proceedings of the IEEE International Conference on Robotics and Automation, Hong Kong, China*, 2014, pp. 1–8.

[16] A. Kahn, J. Marzat, H. Piet-Lahanier, and M. Kieffer, "Cooperative estimation and fleet reconfiguration for multi-agent systems," *Proceedings of the IFAC Workshop on Multivehicle Systems, Genova, Italy*, pp. 11–16, 2015.

[17] C. C. Cheah, S. P. Hou, and J. J. E. Slotine, "Region-based shape control for a swarm of robots," *Automatica*, vol. 45, no. 10, pp. 2406–2411, 2009.

[18] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

A New Method for Omni-RGB+D Camera Rig Calibration and Fusion using Unified Camera Model

Ahmad Zawawi Jamaluddin, Cansen Jiang, Osama Mazhar, Olivier Morel, Ralph Seulin, David Fofi

***Abstract*— The proposed vision system is compact and rigid with two fisheye cameras that provide a 360-degree field of view. Alongside, a high-resolution stereo vision camera is mounted to monitor anterior field of view for precise depth perception of the scene. To effectively calibrate the proposed camera system, we offer a novel camera calibration approach taking the advantages of Unified Camera Model representation. The proposed calibration method outperforms the state-of-the-art methods. Moreover, we proposed more affective algorithm in fusing the two fisheye images into a single unified sphere, which offers seamless stitching results. This new omni-vision rig system is designed to obtain sufficient information to be used on a robot for object detection and recognition. A large scale SLAM and dense 3D reconstruction can be achieved taking the advantage of the large field of view.**

I. INTRODUCTION

At all time, the living creatures have been observed by the scientist and the researcher. Their special abilities have been transformed into the usable form by the aid of technology and finally to be used by human or robot. The artificial systems have been developed either to obtain a better and precise result or to replace a human for dangerous missions[1]. A vision system is one of the important components in this research. The research and development on computer vision are extremely increasing in parallel with the development on robotics technology.[2]. The fabrication of hybrid camera systems with the wide field of view, combine with the computer vision technique makes this research more interesting. The objectives of this research were to fully utilise vision sensor as the useful kit for robotics navigation. This paper presents the novel camera system which can provide a 360° field of view and the depth information concurrently. The system minimizes the use of equipment and image-data and has the ability to acquire sufficient information on the scene. The applications of omnidirectional cameras are mainly for robotics such as localization and mapping[3], robot navigation[4], object tracking[5], visual servoing[6], structure from motion/motion from structure[7][8] and virtual reality[9].

A. Proposed System.

The proposed vision system consists of two CCD cameras mounted with a fisheye lens with each has more than 180° field of view. The fisheye cameras placed back to back so that it cover the whole 360° vertically and horizontally. A high-resolution stereo vision camera placed in front of the rig, so that its baseline is in parallel with the baseline of the fisheye cameras. The stereo vision camera named ZED camera, provides a high-resolution RGB image with the depth information. Fig.1 shows the illustration and the views from of camera rig.

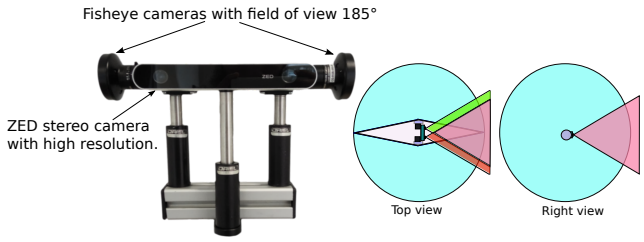


Fig. 1. The front view of the proposed system with the desire field of view.

B. Concept and Motivation

Most of the living creatures have their own capability to look on the object. Some animals have the ability to look from far like an eagle, wide range of view on left and right side like a horse, a motion detector like a fly, thermal view like a snake and stereo and night vision view like a tiger. These abilities of view depend upon the position, form and structure of the eyes. The omni vision cameras rig has been developed by referring to the two categories of animals, prey, and predator. A prey animal has a wide field of view but a small binocular view. They use the eyes for observing the environment from the predator or other threats. While the predator has a large frontal binocular view for targeting and attacking. A creature like jumping spider has the ability to be as prey-like and predator-like. They have four set of eyes for targeting and observing. The same vision system can be realized using multiple camera for the purpose of object observation, recognition and detection.
The major contributions of this paper are three folded:

1) We proposed a very compact Omni-vision system alongside with a Stereo-camera, which offers immense information of 360-degree views of the environment as well as detailed depth information from observation of Stereo-camera.
2) A new camera calibration method taking the advantages of Unified Camera Model representation has been proposed, which outperforms the state-of-the-art methods.
3) To fuse the two fisheye images, an Interior Point Optimization based pure rotation matrix estimation approach has been proposed, which offers seamless image stitching results.

II. HYBRID CAMERA SYSTEM

A. Omnidirectional Camera

There are three main types of the omnidirectional camera which each has their own advantages.

1) Dioptric: A dioptric system or fisheye[10][11], has a field of view more than 180°. It consists of a single camera mounted with a special lens to increase the field of view.

2) Catadioptric: A catadioptric camera is another type of omnidirectional camera[12]. This camera has more than 180° field of view with a single image. It consists of a camera system with a cone shape reflected mirror. This camera produces also a black spot which is the image of the camera itself.

3) Polydioptric: The Greek words 'poly' means many. This camera consist of several overlapping and non-overlapping identical camera to obtain a panoramic 360° field of view.

Fig.2 illustrates the different omnidirectional vision systems.

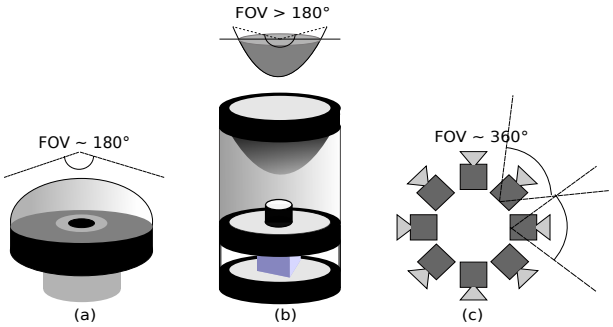


Fig. 2. The illustration shows the omniderectional vision systems (a) Dioptric, (b) Catadioptric and (c) Polydioptric.

The linear perspective geometry has been preserved from any camera system with a single viewpoint. Normally, a single viewpoint exists in all omnidirectional camera. It allows the omnidirectional camera to extract a different view, from perspective to panoramic view[13]. However, for polydioptric which consists of several identical cameras, it considered either they have a unique viewpoint[10][14]. The multi cameras system has a stereovision capability or to enhance the field of view, but it is impossible for the camera rig to have a single effective viewpoint. Fig. 3 shows the illustration of sphere view for the omnidirectional camera.

B. Stereovision Camera

Stereovision is a technique used to build three dimensional description of a scene observed from different viewpoints [15]. If no additional lightening of the scene is required, for example from the laser, this technique is known as passive stereo vision[16]. This technique is used to perceive depth information through generating disparity maps, which then is used to detect obstacles in the environment [17]. This is a classical technique that helps in the field of robotics for

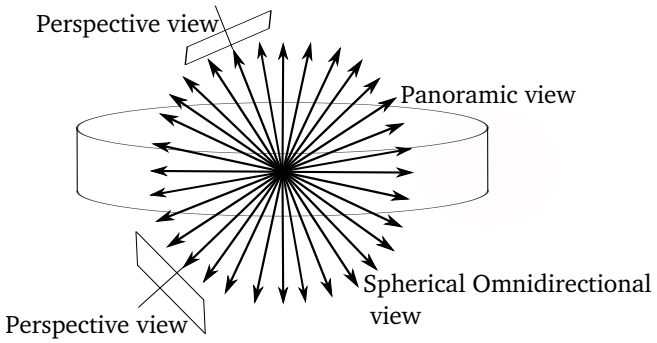


Fig. 3. The illustration shows the sphere view for omnidirectional camera. Image courtesy from Nayar et al: 1997

localization, navigation and obstacle detection since several decades. The development of high-resolution optical sensors and dedicated graphics processing units are helping engineers to design better stereovision cameras for the use in robotics and relevant fields. Lately the release of Bumblebee2 and ZED Stereo Camera[18] enable the researchers to get high-resolution three dimensional depth sensing. Fig. 4 shows the illustration of stereo vison system.

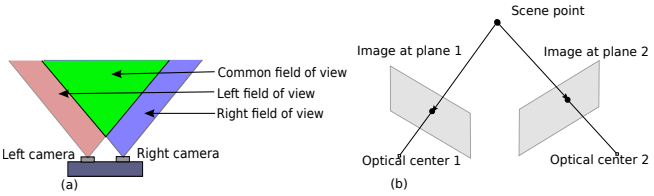


Fig. 4. The Illustration shows the stereo vision system. (a): Left and right cameras have their own view, the green colour is the common view or stereovision. (b): A simple stereovision model.

III. METHODOLOGY

A. Unified Spherical Camera Model

The unified spherical camera model has been proposed by Barreto J.P[14][19]. The image formation on dioptric camera was effected by the radial distortion. Due to that, the point on the scene is not linear with the point on the dioptric image. The image formation in the dioptric camera with radial distortion is in three steps procedure. Following a pinhole camera model, a world point χ originates a projective ray $x = P\chi$, where P is a conventional 3x4 projection matrix. This ray is suppressed into a two dimensional projective point $\mathbf{x} = KP\chi$, where K is the intrinsic parameters matrix. This is the input for the second step, where $\hbar(\cdot)$ (in Equation (10)) is a non-linear transformation in the case of dioptric cameras , where ξ defines the amount of radial distortion.

Then the model was extended by Christopher Mei[20] and he developed the calibration toolbox to calibrate the omni-directional camera. The model was enhanced by introducing another type of distortion called tangential distortion which is incorporated with the radial distortion. Fig. 5 shows the Mei's projection model from camera to unified spherical model. The eccentricity, ξ parameter which defines the amount of

distortion has been explained. Mei's projection model [20] has been used as a reference and it provides procedures to map the image on an unified spherical model.
Let consider a 3D point $\chi = (X,Y,Z)^T$ in the world, and project it to the unit sphere with C_m as a center of the sphere.

$$\chi_s = \frac{\chi}{\|\chi\|} = (X_s, Y_s, Z_s)^T, \tag{1}$$

Then the point χ_s mapped to the new reference frame with the new center C_p

$$(\chi_s)_{F_m} \rightarrow (\chi_s)_{F_p} = (X_s, Y_s, Z_s)^T, \tag{2}$$

Next, the point projected onto the normalised plane

$$m_u = (\frac{X_s}{Z_s+\xi}, \frac{Y_s}{Z_s+\xi}, 1)^T = \hbar X_s, \tag{3}$$

The model of distortion (tangential and radial) are added to the projection model. It consists of three radial and two tangential distortion parameters.

$$x_c = x1 + k_1 r^2 + k_2 r^4 + k_5 r^6 + 2k_3 xy + k_4(r^2+2x^2), \tag{4}$$

$$y_c = y1 + k_1 r^2 + k_2 r^4 + k_5 r^6 + 2k_4 xy + k_3(r^2+2y^2), \tag{5}$$

where:

$$r = \sqrt{x^2+y^2}, \tag{6}$$

and the sum of distortion is

$$m_d = m_u + D(m_u, V), \tag{7}$$

where V contains the coefficients of distortion.

$$V = (k_1, k_2, k_3, k_4, k_5, k_6), \tag{8}$$

and finally, the point m_d is projected to the image plane using K, which is a generalized camera projection matrix. The value f and η should be also generalized to the whole system (camera+lens).

$$p = Km_d = \begin{bmatrix} f_1\eta & f_1\eta\alpha & u_0 \\ 0 & f_2\eta & v_0 \\ 0 & 0 & 1 \end{bmatrix} m_d, \tag{9}$$

where the $[f_1, f_2]^T$ is the focal length, (u_0, v_0) is the principal point and α is the skew factor. Finally, by using the projection model, the point on the normalized camera plane can be lifted to the unit sphere by the following equation:

$$\hbar^{-1}(m_u) = \begin{bmatrix} \frac{\xi+\sqrt{1+(1-\xi^2)(x^2+y^2)}}{x^2+y^2+1} x \\ \frac{\xi+\sqrt{1+(1-\xi^2)(x^2+y^2)}}{x^2+y^2+1} y \\ \frac{\xi+\sqrt{1+(1-\xi^2)(x^2+y^2)}}{x^2+y^2+1} - \xi \end{bmatrix}, \tag{10}$$

B. Camera Calibration using Zero-degree Overlapping Constrain

For the proposed multi-camera setup, there are two 185^o fisheye cameras rigidly attached opposite to each other. Since the field-of-view (FOV) of the fisheye camera has more than 180^o, the proposed setup has an overlapping area between the two fisheye cameras periphery. We are taking the advantages of overlapping FOV of two fisheye cameras, we propose to use a new fisheye Camera calibration using the constrain

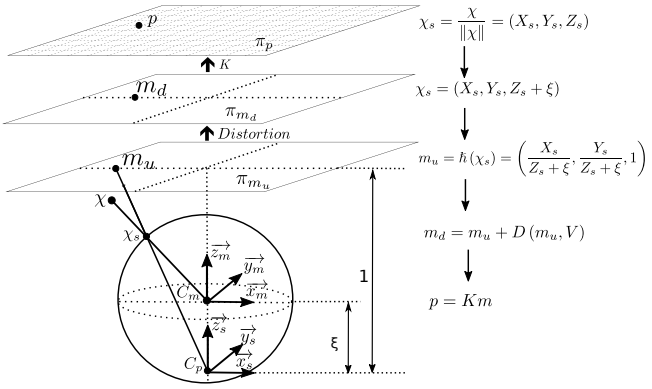


Fig. 5. Mei's projection model from camera to unified spherical model. This figure is the courtesy of Christopher Mei [34].

of overlapping Zero-degree lines of the two fisheye cameras using a Unified Camera Model. Fig. 6 shows experimental setup to re-estimate the value of ξ.

Assumption:

- If ξ is estimated correctly, the 180^o line of the Fisheye camera on the zero degree plane of the Unit Sphere by projecting the fisheye image into a Unified camera model.
- A correct calibration (registration) of multi-fisheye camera setup comes from a correct overlapping area.

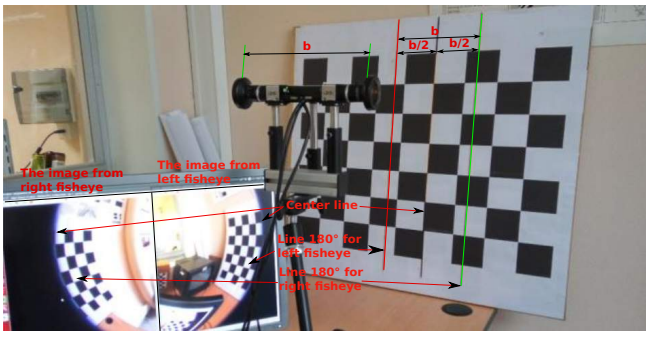


Fig. 6. Experimental setup to calibrate the value of ξ. The baseline of the camera rig (from left fisheye lens to right fisheye lens) is measured and two parallel lines with the same distance to each other as well as a centre line is drawn on a pattern. The rig is faced and aligned in front of the pattern such that the centre line touches the edges of both fisheye camera images

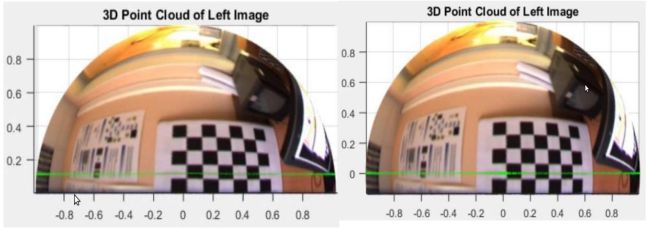


Fig. 7. The left image was projected with initial estimate of ξ; the 180^o lines should ideally lie on the zero plane. After the iterative estimation of ξ, the 180^o line now lie on the zero plane.

1) Zero-crossing Plane Distance Minimization: Let $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\chi_i\}_{i=1}^n$ be a set of points located on the 180^o

line of the fisheye image and their projections onto a unit sphere, receptively. Fig. 7 shows the unit sphere enfolded using initial and estimated ξ and the calibration problem can be simplified as a minimization problem such that the distance between the 180^o-line and the zero-crossing plane of unified sphere is minimized. In other words, the distance between χ and the Zero-crossing plane of the unit sphere is required to be minimized. It denoted as:

$$\min_{\xi} \sum_{i=1}^{n} \psi(\|f(\mathbf{x}_i, \xi)\|_2), \tag{11}$$

where $f(\mathbf{x}, \xi)$ equation (10) is the mapping function from fisheye image ($\mathbf{x}_i$) to the unified camera model (χ_i), operator $\|\cdot\|$ stands for the *l2*-norm, $\psi(\cdot)$ function is the adopted Huber-Loss function for robust estimation purpose. Since the mapping function $f(\mathbf{x}, \xi)$ in equation (11) is not linear, we suggest the Interior Point Optimization Algorithm scheme which is a global non-linear optimization method.

2) Pure Rotation Registration: One of the major objective of our setup is to produce a high quality 360^o FOV unit sphere for handy visualization. To do so, a common way is to calibrate the camera setup such that the relative poses between the cameras are known. Let the features from the left and right fisheye camera (projected onto unit sphere) be denoted as χ^L and χ^R, respectively. The transformation between the two fisheye camera is noted as $T \in R^{4\times4}$, such that:

$$\chi^L = T\chi^R, \tag{12}$$

The transformation matrix, T was estimated, as discussed in [21], the Extrinsic Calibration method using [21] or Singular Value Decomposition (SVD)[24] is not able to correct recover the transformation matrix T due to its pure rotation property. We use a pure rotation matrix to solve this problem by enforcing the transformation matrix contained zero translation[22], which represented as:

$$\min_{R} \sum_{i=1}^{n} \Psi(\|\chi_i^L - R\chi_i^R\|), \quad \text{s.t.} \quad \mathrm{RR}^\mathsf{T} = 1, \det(\mathrm{R}) = 1, \tag{13}$$

Where R is the desired pure rotation matrix, $\Psi(\cdot)$ function is the Huber-Loss function for robust estimation. By solving the above equation, a pure rotation matrix that minimize the registration errors between the fusion of two fisheye cameras. Here, we adopt the Interior Point Optimization algorithm to solve the system.

3) Fusion of Multi-camera Images: In our setup (or other multi-camera setup), the fusion of fisheye cameras alongside with the ZED-camera based on a unified model representation can be achieved in a similar manner. Let $\mathbf{x}^z$ and $\mathbf{x}^{Fl}$ be image feature correspondences between ZED camera and fisheye camera, respectively. Let χ^Z and χ^{Fl} be the feature correspondences (mapped from $\mathbf{x}^z$ and $\mathbf{x}^{Fl}$) on a Unified Sphere. The fusion of the ZED camera and the fisheye camera can be framed as a minimization problem from the feature correspondences on a unified sphere, which can be defined as:

$$\underset{\theta_{x,y,z}}{\operatorname{argmin}} \sum_{i=1}^{n} \Psi\left(\left\|\chi_S^{Lf} - \chi_S^Z(\theta_{x,y,z})\right\|_2\right), \tag{14}$$

where $\Psi(\cdot)$ is the Loss function for the purpose of robust estimation, while

$$\chi(\theta_{x,y,z}) = R(\theta_{x,y,z}) \begin{bmatrix} x_s & . & . & . & . & x_s^n \\ y_s & . & . & . & . & y_s^n \\ z_s & . & . & . & . & z_s^n \end{bmatrix}, \tag{15}$$

stands for the registration of Zed camera sphere points to the left fisheye camera (the reference), where $R(\theta_{x,y,z})$ is the desired pure rotation matrix with estimated rotation angles $\theta_{x,y,z}$. To solve this problem, similar to solving Equation (13), an Interior Point Optimization algorithm is applied.

C. Epipolar Geometry of Omnidirectional Camera.

The epipolar geometry for an omnidirectional camera has been studied and it originally used for a catadioptric camera as a model[17]. The study was extended to the dioptric or fisheye camera system. Fig. 8 shows the epipolar geometry of fisheye camera. Lets consider the two positions of a fisheye camera which observed a point P in the space. Points P_1 and P_2 are the projection of point P onto unit spheres with a coordinate (u_1, v_1) and (u_2, v_2) on the fisheye images[19]. The points P, P_1, P_2, O_1 and O_2 are coplanar,and it can be written as:

$$\begin{aligned} \overline{O_1O_2} \times \overline{O_2P_1} \cdot \overline{O_2P_2} = 0, \\ O_1^2 \times P_1^2 \cdot P_2 = 0, \end{aligned} \tag{16}$$

where, O_1^2 and P_1^2 are the coordinates of O_1 and P_1 in coordinate system X_2, Y_2, Z_2. The transformation between system X_1, Y_1, Z_1 and X_2, Y_2, Z_2 can be described by rotation R and translation t. The transformation equations are:

$$\begin{aligned} O_1^2 = R \cdot O_1 + t = t, \\ P_1^2 = R \cdot O_1 + t, \end{aligned} \tag{17}$$

By substituting (17) in (16) we get,

$$P_2^T E P_1 = 0, \tag{18}$$

where, $E = [t]R$ the essential matrix which consists of rotation and translation. the computation is always to minimizing the epipolar errors. In order to estimate the essential matrix, the points correspondence pairs on the fisheye imsges are stacked into the linear system, thus the overall epipolar constraint becomes.

$$Uf = 0, \tag{19}$$

where,

$$U = [u_1, u_2, ..., u_n]^T \ ,$$

and u_i and f are vectors constructed by stacking column of matrices P_i and E respectively.

$$P_i = P_i P_i'^T, \tag{20}$$

$$E = \begin{bmatrix} f_1 & f_4 & f_7 \\ f_2 & f_5 & f_8 \\ f_3 & f_6 & f_9 \end{bmatrix}, \tag{21}$$

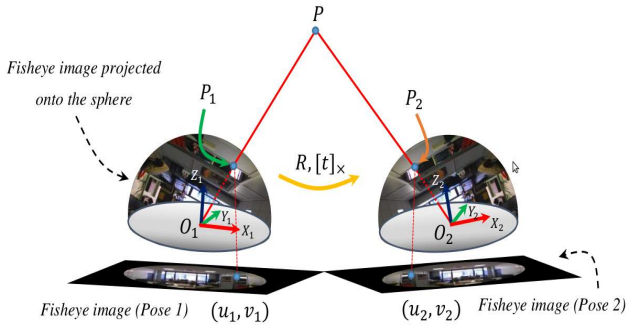


Fig. 8. The epipolar geometry of fisheye camera.

The essential matrix can be estimated with linear least square by solving equation (18) and (19), where P_i^i is the projected point which corresponds to P_2 of the Fig. 6, U is $n \times 9$ matrix and f is 9×1 vector containing the 9 elements of E. The initial estimate of essential matrix is then utilized for the robust estimation of essential matrix and a modified iteratively re-weighted least square method for omnivision cameras is proposed which is originally explained in [24]. This assigns minimal weights to the outliers and noisy correspondences. The weight assignment is performed by finding the residual r_i for each point.

$$r_i = f_1 x_i' x_i + f_4 x_i' y_i + f_7 x_i' z_i + f_2 x_i y_i' \tag{22}$$

$$+ f_5 y_i y' i + f_8 y_i' z_i + f_3 x_i z_i' + f_6 y_i z_i' + f_9 z_i z_i' \ ,$$

$$err \rightarrow \min_f \sum_{i=1}^{n} \left(w_{Si} f^T u_i \right)^2 , \tag{23}$$

$$w_{Si} = \frac{1}{\nabla r_i}, \tag{24}$$

$$\nabla r_i = (r_{xi}^2 + (r_{yi}^2 + (r_{zi}^2 + (r_{xi'}^2 + (r_{yi'}^2 + (r_{zi'}^2)^{\frac{1}{2}}, \tag{25}$$

where, w_{Si} s the weight (known as Sampson's weighting) that will be assigned to each set of corresponding point and ∇r_i is the gradient; r_{xi} and so on are the partial derivatives found from equation (17), as $r_{xi} = f_1 x_i' + f_2 y_i' + f_3 y_i'$.
Once all the weights are computed, U matrix is updated as follow,

$$U = W \times U, \tag{26}$$

where, W is a diagonal matrix of the weights computed using equation (18). The essential matrix is estimated at each step and forced to be of rank 2 in each iteration. The procrustean approach is adopted here and singular value decomposition is used for this purpose.

IV. EXPERIMENTAL RESULTS

A. The estimation of intrinsic parameters

- The unknown parameters f_1, f_2, u_0, v_0 and ξ of fisheye cameras are estimated using the omnidirectional camera toolbox provided by Christopher Mei's.
- The fisheye images are projected onto the unit sphere using the Inverse Mapping Function defined in Christopher Mei's camera projection model.
- Fig. 9 shows the images from left and right fisheye cameras projected onto the unit sphere.

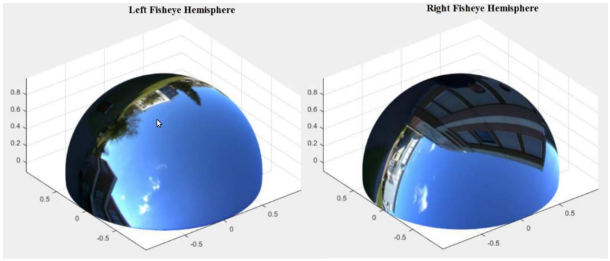


Fig. 9. The images from left and right fisheye cameras projected onto the unit sphere.

1) Re-estimating parameter ξ: The baseline of the camera rig (from left fisheye lens to right fisheye lens) is measured and two parallel lines with the same distance to each other as well as a center line is drawn on a pattern. The rig is faced and aligned in front of the pattern such that the center line touches the edges of both circular fisheye camera images (see Fig. 6 and 7). The parallel line corresponding to the edge of fisheye lens of each camera is then forced to the zero plane when the fisheye image is projected onto the unit sphere. This is done by developing a cost function to estimate ξ that minimizes the z-component of pixels on the selected line using interior point optimization algorithm. This has been explained in the Section III.

B. The estimation of extrinsic parameters

1) Rigid transformation between two fisheyes image: Our system has an overlapping features about 5^o between the left and right hemispheres. The selected points are projected onto the unit sphere. The rigid 3D transformation matrix are estimated using the overlapping features. The Interior Point Optimization Algorithm is used to estimate the rotation between a set of projected points. Fig. 10 shows that the set of projected points are aligned together. The rotation matrix

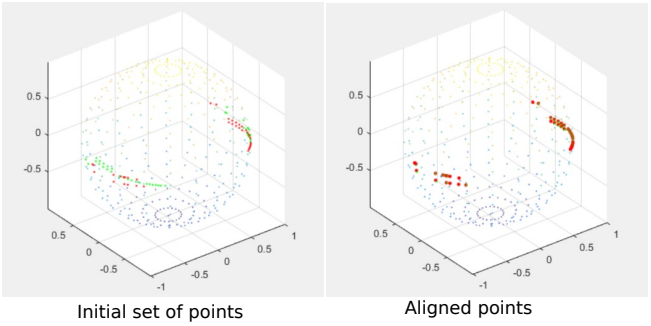


Fig. 10. The selected points are aligned together.

is parameterized in terms of Euler angles and cost function is developed that minimize the Euclidian distance between the reference (point projections of left camera image) and the three dimensional points from the right camera image.

The transformation matrix of two fisheye cameras using our method.,

$$T_{Our} = \begin{bmatrix} -1.0000 & -0.0048 & 0.0085 & 0 \\ -0.0045 & 0.9994 & -0.0335 & 0 \\ -0.0087 & -0.0334 & -0.9994 & 0 \\ 0 & 0 & 0 & 1.0000 \end{bmatrix}, \quad (27)$$

The transformation matrix of two fisheye cameras using Singular Value Decomposition(SVD).

$$T_{SVD} = \begin{bmatrix} -1.0000 & -0.0017 & 0.0073 & -0.0039 \\ -0.0019 & 0.9997 & -0.0250 & 0.0085 \\ -0.0073 & -0.0250 & -0.9997 & 0.1143 \\ 0 & 0 & 0 & 1.0000 \end{bmatrix}, \quad (28)$$

The transformation results using SVD though are very close to pure rotation. It assumed that translation is also as a parameter to align the set of points.

Fig. 11 shows that the two hemispheres are fused together using the estimated transformation matrix. Once the rotation

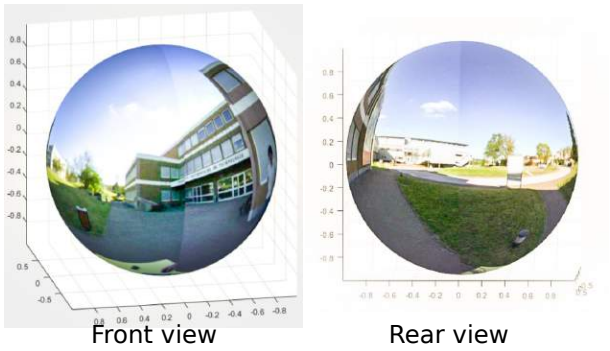


Fig. 11. The unit sphere; images front and rear view from the camera rig.

matrix is estimated, fusion of the two hemispheres is only two step procedure. The points on the hemispheres that are beyond the zero plane are first eliminated. Then the transformation (rotation only) is applied on the hemisphere of the right fisheye camera and the point matrices are concatenated to get a full unit sphere.

2) Rigid transformation between a fisheyes and ZED camera: The same procedures are used to estimate the transformation matrix between the image from ZED camera and two hemispheres. The transformation matrix image ZED camera refers to left fisheye.

$$T_{ztolf} = \begin{bmatrix} -0.0143 & -0.0290 & -0.9995 & 0 \\ -0.0062 & 0.9996 & -0.0289 & 0 \\ -0.9999 & 0.0058 & -0.0145 & 0 \\ 0 & 0 & 0 & 1.0000 \end{bmatrix}, \quad (29)$$

As shown in fig. 12, the RGB and depth images from ZED camera are overlapped onto the unit sphere. It is also recovered the scale between the ZED and fisheye cameras. The high-resolution RGB image and the depth image are fused onto the unit sphere by the help of transformation matrix estimated. It should be noted that, the fusion of zed image onto the unit sphere is an approximation. The result obtained is reasonable after handling the strong distortion on the fisheye images.

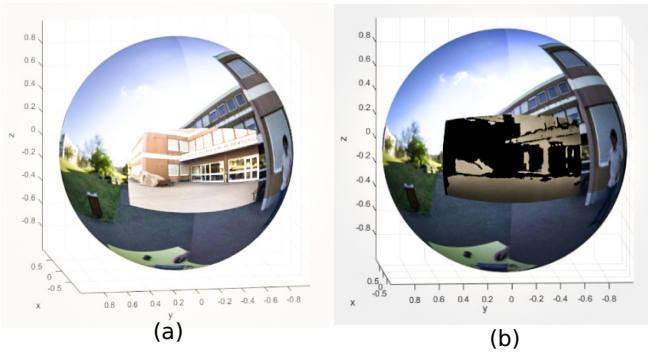


Fig. 12. The unit sphere overlaps with image from ZED camera (a: RGB image, b: Depth image).

C. Estimation the three dimensional registration error.

The computation of registration error during mapping on the unit sphere is done to proof the registration method. The Root Means Square Error (RMSE) is frequently used to calculate the error. The rigid 3D transformation matrix and parameter ξ which was obtained from calibration are used to determine the residuals error which is the difference between the actual values and the predicted values. Three methods have been compared:

1) IPOA : The pure rotation estimated using feature matches and Interior Point Optimization Algorithm.
2) SVD : The transformation (rotation and translation) estimated using features matches with Singular Vector Decomposition[24].
3) CNOC : The method used in Calibration Non Overlapping Cameras[21].

The image sequences were taken in several different environments. The feature points were selected on the overlapping area. The same data set are used in all three methods. Fig. 13, Fig. 14 and Fig. 15 show that the proposed method has the lowest registration errors.

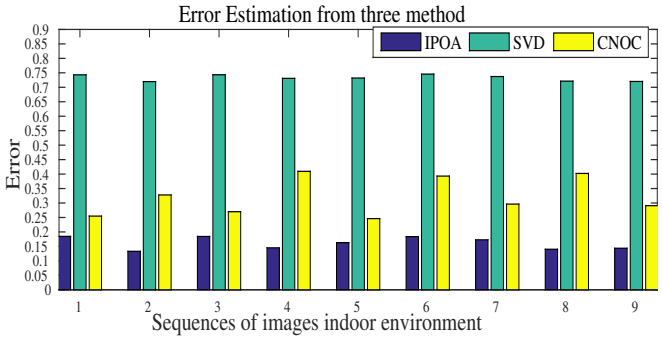


Fig. 13. The registration error estimation using three different methods. For the first experiment, the images sequences have been taken inside the building. The average registration error using proposed method is 0.1612.

V. CONCLUSIONS

The intrinsic parameter of fisheye and ZED cameras are estimated from the camera calibration toolbox. The distortion of fisheye camera is determined by parameter ξ. We were

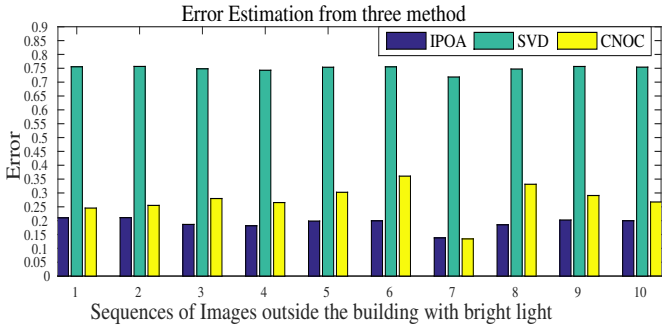


Fig. 14. For the second experiment, the images sequences have been taken outside the building with bright light. The average registration error for the proposed method is 0.1912.

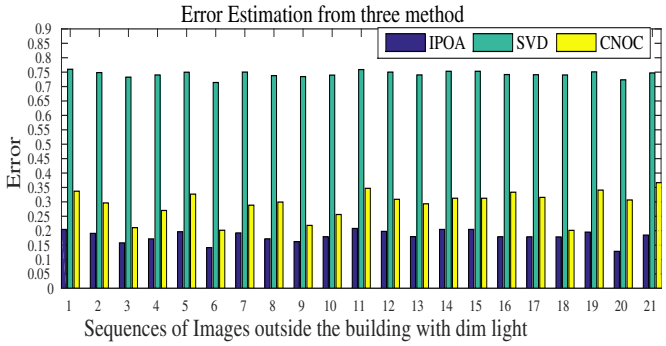


Fig. 15. For the third experiment, the images sequences have been taken outside the building with dim environment (cloudy). The average registration error using proposed method is 0.1812

re-estimating the value of parameter ξ by using a cost function to minimize the z-component of pixels on the selected line using interior point algorithm. The rigid 3D transformation matrix are estimated using the overlapping features assuming that it is a pure rotation. The Interior Point Optimization Algorithm is used to estimate the rotation between a set of projected points. The same procedure is used to fuse ZED camera onto the unit sphere. The overlap area between fisheye and ZED cameras is estimated for the purpose of object tracking and detection. The camera rig is applied to reconstruct 3D scene using features matching and an algorithm based on a spherical model of camera. The registration error is calculated to show the performance of our proposed method.

REFERENCES

[1] Marefat, F., A. Partovi, and A. Mousavinia. "A hemispherical omni-directional bio inspired optical sensor." 20th Iranian Conference on Electrical Engineering (ICEE2012). IEEE, 2012.

[2] Li, Shigang. "Full-view spherical image camera." 18th International Conference on Pattern Recognition (ICPR'06). Vol. 4. IEEE, 2006.

[3] Liu, Ming, and Roland Siegwart. "Topological mapping and scene recognition with lightweight color descriptors for an omnidirectional camera." IEEE Transactions on Robotics 30.2 (2014): 310-324.

[4] Zhang, Chi, et al. "Development of an omni-directional 3D camera for robot navigation." 2012 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, 2012.

[5] Marković, Ivan, Franois Chaumette, and Ivan Petrovi. "Moving object detection, tracking and following using an omnidirectional camera on a mobile robot." 2014 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2014.

[6] Depraz, Florian, et al. "Real-time object detection and tracking in omni-directional surveillance using GPU." SPIE Security+ Defence. International Society for Optics and Photonics, 2015.

[7] Chang, Peng, and Martial Hebert. "Omni-directional structure from motion." Omnidirectional Vision, 2000. Proceedings. IEEE Workshop on. IEEE, 2000.

[8] Micusik, Branislav, and Tomas Pajdla. "Structure from motion with wide circular field of view cameras." IEEE Transactions on Pattern Analysis and Machine Intelligence 28.7 (2006): 1135-1149.

[9] Li, Dong, et al. "Motion Interactive System with Omni-Directional Display." Virtual Reality and Visualization (ICVRV), 2013 International Conference on. IEEE, 2013.

[10] Nayar, Shree K. "Catadioptric omnidirectional camera." Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on. IEEE, 1997.

[11] Neumann, Jan, Cornelia Fermuller, and Yiannis Aloimonos. "Polydioptric camera design and 3d motion estimation." Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. Vol. 2. IEEE, 2003.

[12] Gluckman, Joshua, and Shree K. Nayar. "Ego-motion and omnidirectional cameras." Computer Vision, 1998. Sixth International Conference on. IEEE, 1998.

[13] Knill, Oliver, and Jose Ramirez-Herran. "Space and camera path reconstruction for omni-directional vision." arXiv preprint arXiv:0708.2442 (2007).

[14] Geyer, Christopher, and Kostas Daniilidis. "A unifying theory for central panoramic systems and practical implications." European conference on computer vision. Springer Berlin Heidelberg, 2000.

[15] Hartley, Richard I., and Peter Sturm. "Triangulation." Computer vision and image understanding 68.2 (1997): 146-157.

[16] Ayache, Nicholas, and Francis Lustman. "Trinocular stereovision for robotics." IEEE Transactions on Pattern Analysis and Machine Intelligence 13.1 (1991).

[17] ZED Stereo Camera from https://www.stereolabs.com/zed/specs/.

[18] Kumar, Saurav, Daya Gupta, and Sakshi Yadav. "Sensor fusion of laser and stereo vision camera for depth estimation and obstacle avoidance." International Journal of Computer Applications 1.25 (2010): 20-25

[19] Barreto, João P. "A unifying geometric representation for central projection systems." Computer Vision and Image Understanding 103.3 (2006): 208-217.

[20] Mei, Christopher, and Patrick Rives. "Single view point omnidirectional camera calibration from planar grids." Proceedings 2007 IEEE International Conference on Robotics and Automation. IEEE, 2007.

[21] Lébraly, Pierre, et al. "Calibration of non-overlapping cameras-application to vision-based robotics." (2010).

[22] Othmani, Alice Ahlem, et al. "A novel Computer-Aided Tree Species Identification method based on Burst Wind Segmentation of 3D bark textures." Machine Vision and Applications (2015): 1-16.

[23] Harris, Chris, and Mike Stephens. "A combined corner and edge detector." Alvey vision conference. Vol. 15. 1988.

[24] ourakis, Manolis IA, and Rachid Deriche. Camera self-calibration using the singular value decomposition of the fundamental matrix: From point correspondences to 3D measurements. Diss. INRIA, 1999.

Remote Photoplethysmography with Constrained ICA using Autocorrelation as a periodicity measure

Richard Macwan, Yannick Benezeth, Alamin Mansouri
Le2i UMR6306, CNRS, Arts et Métiers
Univ. Bourgogne Franche-Comté

Abstract

Remote photoplethysmography(rPPG) is being increasingly used to measure heart rate from recorded or live videos. The rhythmic flow of arterial blood, referred to as the blood volume pulse, results in periodic variations in the skin color which are then quantified into a temporal signal for analysis.

We present a novel method for measuring remote photoplethysmography(rPPG) signals using Constrained Independent Component Analysis(cICA). We provide autocorrelation as an a priori information for cICA to extract the most prominent rppg signal. CICA with autocorrelation showed improved performance over traditional ICA in terms of convergence speed and accuracy with two different in-house video databases.

1. Introduction

Photoelectric plethysmography or photoplethysmography(PPG) was first introduced in 1937 by Hertzman where the variations in the light absorption of human skin were measured by a photoelectric cell[7] placed under a finger illuminated by a light source placed above it. Since then PPG has been used widely because of its ease of usage,low cost and non-invasiveness. This non-invasiveness has, however, been superseded by that of remote photoplethysmography, henceforth referred to as rPPG, which aims at measuring the same parameters, but *sans contact*.

Verkrussysse [18] demonstrated the extraction of remote PPG signals using videos from a simple consumer level camera and that the strongest photoplethysmographic signal was manifested in the G channel of the RGB temporal traces. RGB temporal traces are generated by frame wise quantification, e.g. spatial averaging of skin pixels from the face, and concatenating them. Current research focuses on extracting robust rPPG signals from simple web cameras for which Blind Source Separation(BSS) using Independent Component Analysis(ICA) has been used in many different works [4, 14, 15, 16].

ICA is a statistical technique for decomposing a multivariate signal into constituent signals assuming that the input signals are uncorrelated [9]. The problem of rPPG measurements is posed as a signal separation problem where the rhythmic cardiac pulse, appearing as variations in skin color, is assumed to be linearly separated and mixed into the temporal traces of color data from cameras. If the time varying color traces for different channels are represented as $\mathbf{s} = (s_1, s_2, ..., s_n)^T$, which is an instantaneous linear mixture of the original independent signals denoted as $\mathbf{c} = (c_1, c_2, ..., c_m)^T$, then the process of mixing can be formulated as $\mathbf{s} = \mathbf{Ac}$, where the mixing matrix $\mathbf{A}_{n\times m}$ represents the linear memoryless mixing of the channels. The goal of ICA, then, is to estimate the demixing matrix $\mathbf{W_{m\times n}}$ to recover all the ICs from the observed signal with minimal knowledge of $\mathbf{A}$ and $\mathbf{c}$. The recovered signal $\mathbf{y} = (y_1, y_2, ..., y_m)^T$, is given by $\mathbf{y} = \mathbf{Wx}$ [13].

This formulation of ICA is based on linear mixing, and thus suffers from two unavoidable ambiguities [2, 6]. First, the order of the independent components is indeterminable. A different permutation of the columns of $\mathbf{W}$ will give the same independent components. Second, the exact amplitude and sign of the independent components is also indeterminable. In spite of these limitations, ICA is being frequently used for rPPG measurements.

Furthermore, we know that the most periodic signal that might be embedded in the RGB temporal traces must correspond to the blood volume pulse. Additionally, we only require one component, viz., the rPPG pulse from the mixture of the temporal traces. This requirement is quite common in various applications, for example, the On-Off simulation scheme of an fMRI experiments [17]. Consequently, it would be advantageous to provide this a priori knowledge to the component extraction algorithm.

The purpose of Constrained Independent Component Analysis(cICA) is just this, to provide a systematic and flexible method to incorporate more assumptions and prior information into the contrast function to make the ICA problem a better-posed problem. Specifically, cICA avoids the ambiguities of ICA by directly converging to the best independent component. In this paper, we use the periodicity of the rPPG signal as an a priori information to determine the component that is the most periodic one. The optimization is steered in the direction of choosing the component with the highest periodicity which represents the actual blood volume pulse. The cICA algo-

rithm is detailed in section 3.2.

2. Previous Work

One of the first works that used ICA for rPPG measurements comprised of using RGB temporal traces from a simple web camera to extract the cardiac pulse, albeit with limited success under the presence of movement artifacts. Usage of more color channels using a five band camera (RGBCO) with ICA was also investigated [14]. ICA's known caveat of the indeterminacy of the order of the estimated components calls for a heuristic to choose the correct component. Where Poh et. al. [15] simply selected the second obtained component after applying ICA on RGB temporal traces , McDuff et. al. [14] chose the signal with the peak of greatest power of the normalized FFT spectrum between 40 and 180 bpm.

De Haan et. al. [4] introduced chrominance-based methods where two orthogonal chrominance signals were built from the RGB traces in addition to using skin-tone standardization to compensate for illumination variation of different skin colors. They further improved upon the chrominance based methods to show that the different absorption spectra of arterial blood happen along a specific vector in a normalized RGB space, termed as the Blood Volume Pulse vector [5].

In a related work, Wang et. al. [19] attempted to extracted the rPPG signal by constructing pixel based rPPG sensors to estimate a robust pulse signal using motion compensated pixel-to-pixel pulse extraction based on optical flow vectors.

Incorporating a priori information to guide the optimization process is an interesting approach in signal separation. Lu. et. al. [11] have used an existing reference signal to guide the separation process by using the method of Lagrange multipliers where the distance between the reference signal and the estimated signal is taken as the constraint to be minimized. However, such a reference signal is not always present, or at times is difficult to synthesize. In rPPG measurements, a PPG signal is such an example the synthesis of which depends critically on the required frequency, more so than on the actual shape of the signal.

To overcome this limitation, we use autocorrelation as the a priori information for guiding the cICA separation algorithm which then chooses the most periodic component representing the blood volume pulse. The algorithm with the workflow of the whole experiment is presented in the next section.

3. Methods

The workflow of the experimental procedure as depicted in figure 1 is presented here. Temporal RGB traces, $\mathbf{x} = [x_1\, x_2\, x_3]^T$ where each x_m, $m \in [1...3]$, corresponds to a temporal trace of size N of each channel, generated by spatial averaging of the pixels (either entire image, face-cropped or skin-segmented) were fed to the cICA algorithm.

3.1. Preprocessing for cICA

The initial preprocessing steps required for constrained ICA are the same as those for ICA which are generally recommended for simplifying the calculations for obtaining the independent components. After normalizing the RGB traces, *centering* was performed so that the obtained signal $\mathbf{y}$ in $\mathbf{y} = \mathbf{W}\mathbf{x}$ is zero-mean. After centering, *whitening* was performed to ensure that the components were uncorrelated and their variances equal to unity.

The RGB traces were then detrended using a smoothness priors approach proposed by Karjalainen et. al. [10] to remove low frequency trends in the signal. Finally, a fourth order butterworth bandpass filter was used to filter the signal within limits of human heartrate between .3 and 3 Hz. The RGB traces are now ready to be fed into the constrained ICA algorithm for obtaining the rPPG pulse signal.

3.2. RPPG Pulse Extraction using cICA

Biomedical signals like ECG and PPG signals from finger sensors are typically known to be periodic or quasi-periodic. In this work, instead of using a reference signal typically used with the cICA algorithm [11], we exploit the inherent periodicity property of these biomedical signals, thereby guiding the separation process to choose the component with the highest periodicity. We use autocorrelation as a periodicity measure which, owing to the formulation of lagrange multipliers, requires the calculation of the first and second derivatives of the autocorrelation with respect to the weighting matrix $\mathbf{w}$.

3.2.1. Autocorrelation as a periodicity measure

Autocorrelation is the correlation of a signal with itself at different lag times provided it is sampled at a sufficiently high frequency. For a time series signal $\mathbf{y} = [y_1\, y_2 ... y_N]$ of N elements, it's discrete autocorrelation r_k at lags $k \in [-(N-1), \cdots, N-1]$ is given by

$$r_k = \sum_{j=0}^{N-1} y_j \odot \overset{k}{y_j} \tag{1}$$

where $\overset{k}{y_j}$ is the j^{th} element of the signal $\mathbf{y}$ lagged(or led if $k < 0$) by k units and padded with zeroes to the left (or right if $k < 0$) and $\odot$ is the element-wise multiplication operator. A periodic signal typically has a higher correlation with itself compared to a non-periodic one. This high correlation can be quantified as the mean of the squared autocorrelation of the signal and consequently can be used as a measure of the periodicity of a signal. Figure 2 depicts the high correlation of a periodic sinusoid compared to that of a uniform random signal with the mean of

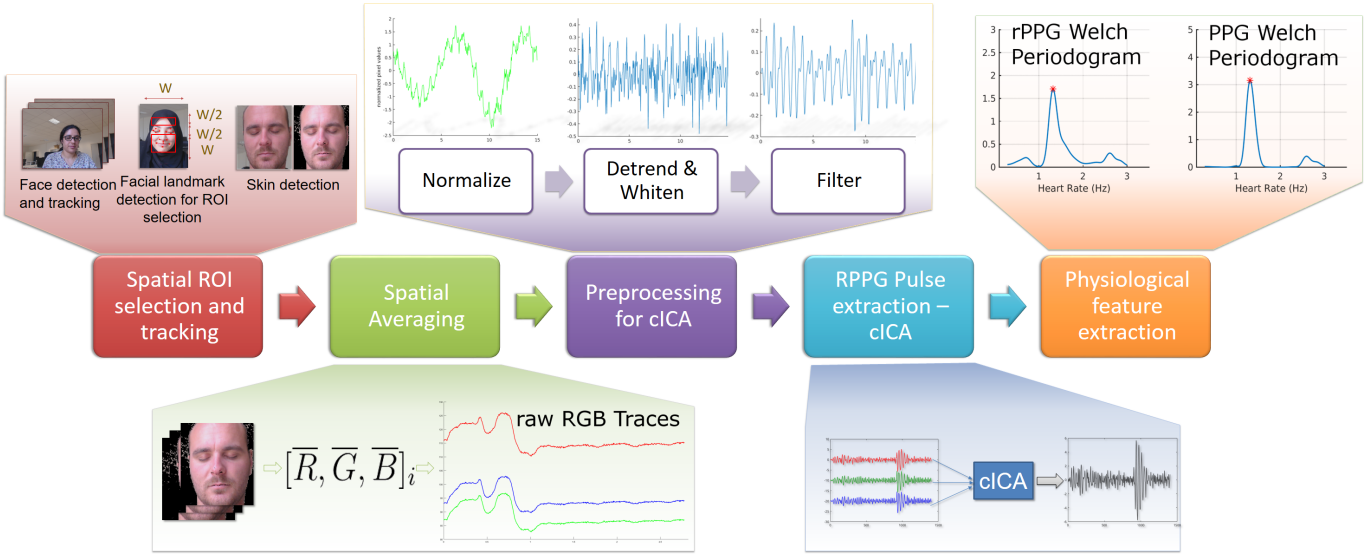
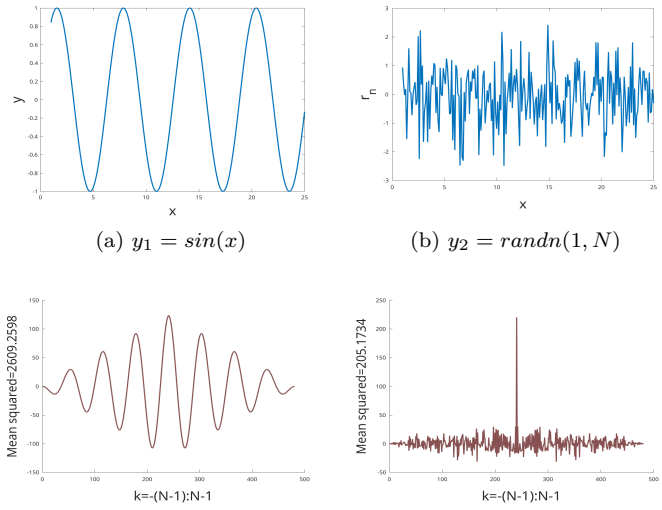


Figure 1. Methodology

the squared autocorrelation is much higher than that of the random signal.



(a) $y_1 = sin(x)$ (b) $y_2 = randn(1, N)$

(c) Autocorrelation of y_1 (d) Autocorrelation of y_2

Figure 2. Autocorrelation of a sinusoid vs a random signal

To aid the use of autocorrelation as a periodicity measure and simplify its computation, two modifications need to be made. First, since the autocorrelation is symmetric, we only compute the correlation for lags $k \in [0, \cdots, N-1]$. Second, since the correlation at lag 0 will always be high, we set the autocorrelation to *0* at lag $k = 0$. Thus, the autocorrelation is given by $\mathbf{r} = [r_1\, r_2 ... \, r_N]$ comprising of N values given by equation 1 and $r_1 = 0$. Keeping in mind that r_k is a scalar, equation 1 can be rewritten in matrix notation as

$$r_k = \mathbf{y}[\overset{k}{\mathbf{y}}]^T = \overset{k}{\mathbf{y}}\mathbf{y}^T \tag{2}$$

where $\overset{k}{\mathbf{y}}$ is again the signal $\mathbf{y}$ lagged by k units. Furthermore, to simplify the derivation of the autocorrelation, $\overset{k}{\mathbf{y}}$ can be rewritten as yT_k where T_k is a toeplitz-like matrix that incorporates the lagging at lag k and padding with zeroes of the signal and is given by

$$T_k = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} = \begin{bmatrix} 0_{N-k,k} & I_{N-k} \\ 0_{k,k} & 0_{k,N-k} \end{bmatrix} \tag{3}$$

T_k is an $N \times N$ matrix composed of the first $N-k$ rows made up of $(N-k) \times k$ zeroes and an identity matrix of size $N-k$. Thus, r_k becomes

$$r_k = \mathbf{y}T_k\mathbf{y}^T \tag{4}$$

making its differential with respect to $\mathbf{y}$ easier to calculate. This autocorrelation is then used to guide the optimization process of constrained ICA to converge to the component with the highest periodicity, which is presented in the next subsection.

3.2.2. Constrained ICA

A generic contrast function for ICA as defined by [9], is the negentropy function given by $J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y})$ where $H(.)$ is the differential entropy and $\mathbf{y}_{gauss}$is a random variable with a variance equal to that of the output signal $\mathbf{y}$. In FastICA, an approximation of the negentropy was introduced for more reliability and flexibility given by

$$J(\mathbf{y}) \approx \rho[E\{G(\mathbf{y})\} - E\{G(v)\}]^2 \tag{5}$$

where ρ is a positive constant, v is a zero mean, unit variance Gaussian and G(.) can be any non-quadratic function. As suggested by [8] a good general purpose function is given by

$$G(y) = \frac{\log\cos(a_1 y)}{a_1} \tag{6}$$

with $1 < a_1 < 2$.

Constrained ICA aims to alleviate the issues of ICA with the help of Lagrange multiplier methods [11]. Lagrange multiplier methods [1] are a tool for performing constrained optimization problems following the general form

$$\text{minimize } f(\mathbf{X}), \text{ subject to } g(\mathbf{X}) \leq 0,\ h(\mathbf{X}) = 0 \quad (7)$$

where $f(\mathbf{X})$ is the objective function, $g(\mathbf{X})$ is a set of inequality constraints and $h(\mathbf{X})$ is a set of equality constraints.

The objective of obtaining the most periodic component using cICA can be fulfilled with the help of the inequality constraint

$$g(\mathbf{w}) = \epsilon(\mathbf{w}) - \zeta \leq 0 \quad (8)$$

where $\mathbf{w}$ represents a single demixing weight vector of size equal to the number of input channels. The optimum $\mathbf{w}$ then extracts the most periodic component using $\mathbf{y} = \mathbf{w}^{\mathrm{T}}\mathbf{x}$. $\epsilon(\mathbf{w})$ represents the closeness measure. Using average of squared autocorrelation as a closeness measure gives $g(\mathbf{w})$ as

$$g(\mathbf{w}) = \zeta - E\{\mathbf{r}^2\} \leq 0 \quad (9)$$

where ζ now denotes the threshold for the lower bound of the optimum autocorrelation. The details of the use of this constraint are presented in the next subsection.

The Augmented Lagrangian for cICA The general cICA problem is defined as [11]

$$\begin{aligned} \text{Maximize}: \quad J(\mathbf{y}) &= \rho[E\{G(\mathbf{w}^{\mathrm{T}}\mathbf{x})\} - E\{G(v)\}]^2, \\ \text{Subject to}: \quad g(\mathbf{w}) &\leq 0,\ h(\mathbf{w}) = E\{\mathbf{y^2}\} - 1 = 0 \end{aligned} \quad (10)$$

where $J(\mathbf{y})$ is the one-unit contrast function as defined in equation 5, $g(\mathbf{w})$ is the closeness constraint and $h(\mathbf{w})$ constrains the output $\mathbf{y}$ to have unit variance.

Finally, the augmented Lagrangian method was used owing to its wider applicability and improved stability [1]. The augmented Lagrangian for our formulation as adapted from [11] is given by

$$\begin{aligned} \mathcal{L}_1(\mathbf{w}, \mu, \lambda) &= J(\mathbf{y}) - \frac{1}{2\gamma}\left[\{[max\{0, \bar{g}(\mathbf{w})\}]^2 - \mu_i^2\}\right] \\ &\quad -\lambda h(\mathbf{w}) + \frac{1}{2}\gamma_1 \left\|h(\mathbf{w})\right\|^2 \end{aligned} \quad (11)$$

where $\bar{g}(\mathbf{w}) = \mu + \gamma g(\mathbf{w})$, μ and λ are the lagrange multipliers corresponding to $g(\mathbf{w})$ and $h(\mathbf{w})$ respectively.$\|.\|$ denotes the Euclidean norm and the term $\frac{1}{2}\gamma \left\|.\right\|^2$ is the penalty term that makes sure that the optimization problem is held at the condition of local convexity assumption:$\nabla^2_{xx}\mathcal{L} > 0$.

To find the maximum of $\mathcal{L}_1$ in equation 11 a Newton-like learning method can be used to iteratively adapt $\mathbf{w}$

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta(\mathcal{L}^{''}_{1_{\mathbf{w}_k^2}})^{-1}\mathcal{L}^{'}_{1_{\mathbf{w}_k}} \quad (12)$$

where k is the iteration index, η is the positive learning rate to avoid uncertainty in convergence and $\mathcal{L}^{'}_{1w}$ is the first derivative of $\mathcal{L}$ w.r.t $\mathbf{w}$ given by

$$\mathcal{L}^{'}_{1w} = \bar{\rho}E\{\mathbf{x}G^{'}_y(\mathbf{y})\} - \frac{1}{2}\mu E\{g^{'}(\mathbf{w})\} - \lambda E\{\mathbf{xy}\} \quad (13)$$

where $\bar{\rho} = \pm\rho$ depending on the sign of $E\{G(\mathbf{y})\} - E\{G(v)\}$, $G^{'}_y(\mathbf{y})$ and $g^{'}(\mathbf{w})$ are the first derivatives of $G(\mathbf{y})$ and $g(\mathbf{w})$ w.r.t $\mathbf{y}$ and $\mathbf{w}$ respectively. The Hessian $\mathcal{L}^{''}_{1_{\mathbf{w}_k^2}}$ in equation 12, is calculated as

$$\mathcal{L}^{''}_{1_{\mathbf{w}_k^2}} = \bar{\rho}\mathbf{R_{xx}}E\{G^{''}_y(\mathbf{y})\} - \frac{1}{2}\mu E\{g^{''}(\mathbf{w})\} - \lambda \quad (14)$$

the inversion of which is not problematic because $\mathbf{R_{xx}}$ being the covariance matrix of the whitened and centered signal $\mathbf{x}$ is an identity matrix. $G^{''}_{y^2}(\mathbf{y})$ and $g^{''}(\mathbf{w})$ are second order derivatives and $\mathcal{L}_{1_{w_k^2}}$ is of size $m \times m$. The first and second derivatives of $g(\mathbf{w})$ are not trivial and are presented in the appendix. Finally, the approximate Newton learning step is given by

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta\mathcal{L}^{'}_{1_{\mathbf{w}_k}}/\mathcal{L}^{''}_{1_{\mathbf{w}_k^2}} \quad (15)$$

The optimum multipliers μ_i^* and λ^* are obtained iteratively based on a gradient-ascent method [12]:

$$\mu_{k+1} = max\{0, \mu_k + \gamma g(\mathbf{w}_k)\}, \quad (16)$$

$$\lambda_{k+1} = \lambda_k + \gamma h(\mathbf{w}_k) \quad (17)$$

Following the above equations, the optimization procedure converges to the optimum point defined by the triplet (w^*, μ^*, λ^*) representing the tuned parameters and final weighting matrix $\mathbf{w}*$ which is then used to obtain the final rPPG signal.

3.3. Physiological feature extraction

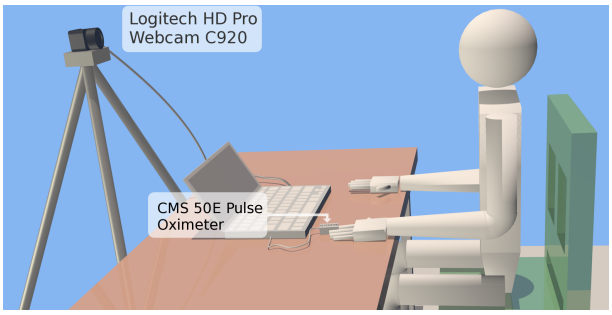
After the rPPG signal was obtained, the per window heart rate was calculated from the highest peak of the periodogram obtained using Welch's method. over a 15 second moving window using a step size of 1 second. ICA (and consequently cICA) is known to work much better with a signal of longer duration. However, to emulate a live scenario as closely as possible, steps 3.1 to 3.3 were performed for each window, using the weighting matrix $\mathbf{w}_k$ at step k as an initial estimate for calculation of $\mathbf{w}_{k+1}$ at the next step. Kalman filtering was used for predicting HRs for windows with $\delta = |HR_{rPPG} - HR_{PPG}| > 10\,bpm$. We present the results of the experiments of windowed and entire signal analysis in the next section.

4. Results and Discussion

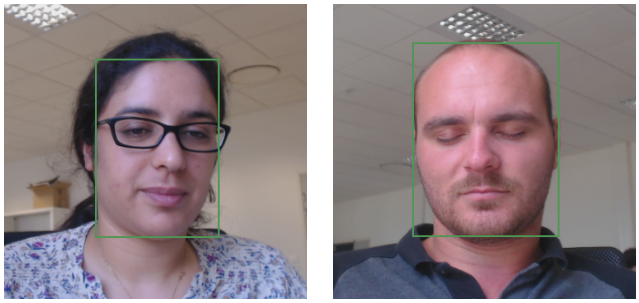
An in-house dataset comprising of 9 videos of around 90 seconds each was used to test the performance of the cICA method where the subjects were asked to sit in a relaxed pose in front of the camera. The video frames were obtained with a custom C++ application using a Logitech C920 web camera placed at a

distance of about 1m from the subject with a resolution of 640x480 in 8-bit uncompressed RGB format at 30 frames per second. A CMS50E transmissive pulse oximeter was used to obtain the ground truth PPG data. The experimental setup with images from two different videos showing the light conditions is depicted in figure 3.

To generate the temporal RGB traces, face detection was first performed using the Viola-Jones implementation provided by the computer vision toolbox of MATLAB. Corner detection in the detected face was performed for tracking to select the facial landmarks. Skin detection as formulated by Conaire et. al. [3] was then performed to select the candidate pixels which were then spatially averaged to obtain a triplet of RGB values per frame which were then concatenated to obtain the final RGB temporal traces.



(a) Experimental seup



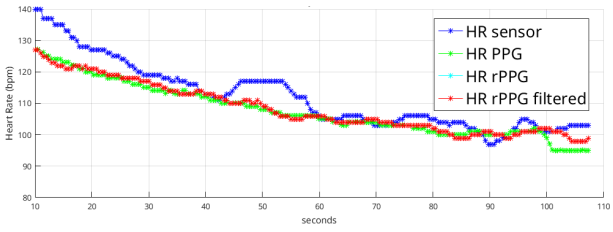
(b) Images from two videos

Figure 3. Experimental Setup

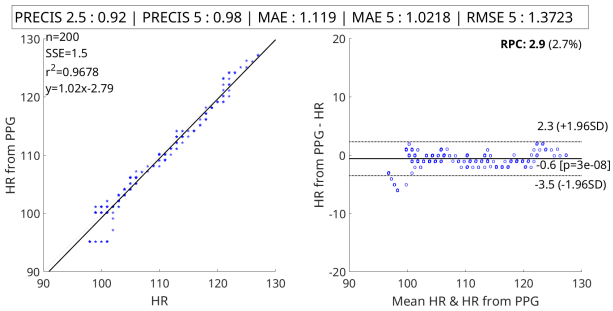
Figure 5c shows a typical rPPG signal and HR_{rPPG} vs HR_{PPG}. The slight phase shift in between the peaks of rPPG and PPG can be attributed to the pulse transit time: the time difference between the cardiac pulse wave reaching different peripheral organs. Figure 5a shows the Bland Altman (BA) analysis performed for the same video on each window. The window-wise difference between HR_rPPG and HR_PPG varies between +1.7 to -2 bpm and has a mean absolute error, $MAE = .63$.

One of the videos in the dataset was recorded after performing physical exercise to verify the drop in HR over time. The cICA method performs well in this case. The temporal comparison of HR from rPPG vs HR from PPG and its corresponding BA analysis is presented in figure 4.

Table 1 shows the accuracy comparisons between ICA and cICA using the mean absolute error (MAE) between HR_{rPPG} and HR_{PPG}. The windowed



(a) RPPG



(b) Bland Altman Analysis

Figure 4. cICA for after-exercise video

	Full	Face	Skin	Full	Face	Skin
	Entire signal			Windowed		
ICA	5.34	6.53	2.1	14.81	13.25	14.12
cICA	0.82	0.79	0.81	8.91	8.09	5.62

Table 1. Mean Absolute Error of ICA vs cICA

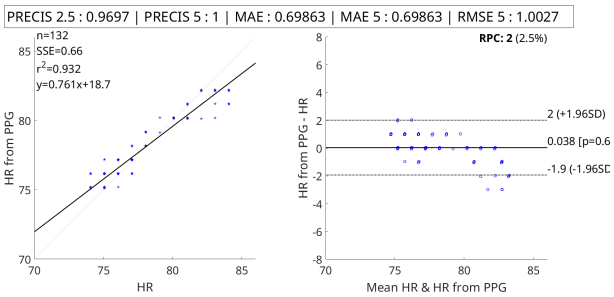
method is computationally more taxing but is more realistic and at understandably less accurate since we are operating only on a 15 second window of the signal which makes the component separation more difficult.

A global Bland Altman analysis was also performed using window-wise calculations from all the videos in each dataset for three different configurations: the entire image, face-cropped image and skin-segmented image. Figure 5b shows the BA analysis for the entire dataset using the entire image. The metrics PRECIS 2.5 and PRECIS 5 show the percentage of windows where $\delta = (HR_{rPPG} - HR_{PPG}) < 2.5$ and 5 bpm respectively. Correspondingly, the MAE 5 and the RMSE 5 metrics measure the mean absolute difference and the root mean squared difference respectively for windows with $\delta < 5$ bpm.

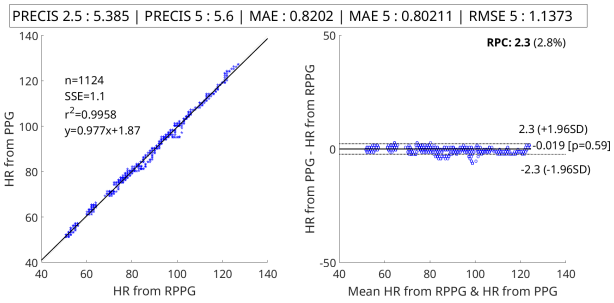
5. Conclusions and future work

In this paper we presented a novel semi blind source separation method for the application of rPPG measurements using autocorrelation as the constraint to guide the ICA separation process. The cICA using autocorrelation provides better result than simple ICA while removing the extra step for choosing the best component. The periodogram of the extracted signals was also constintently closer to that of the PPG.

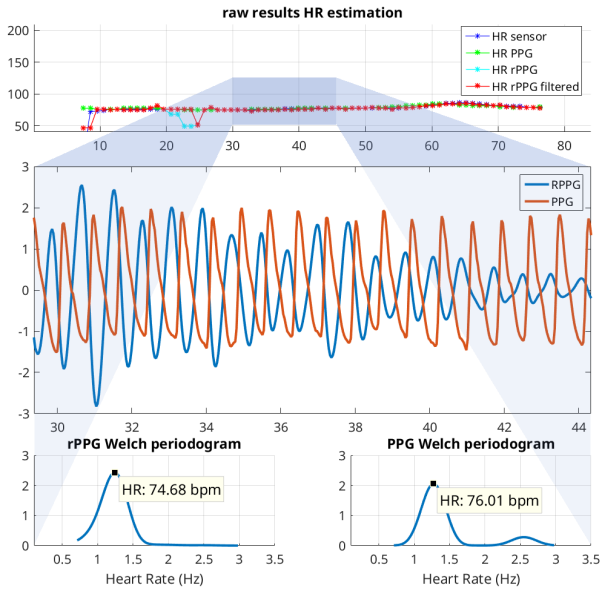
For improving accuracy, better face and skin detectors and trackers can be investigated. Also, the assumption that the most periodic component is the cardiac pulse signal does not hold in scenarios with periodic motion e.g. in fitness. The method can thus



(a) Bland Altman for video 5



(b) Bland Altman for dataset SIMPLE



(c) Extracted rPPG signal for a 15 sec window of video 5. HR_sensor is HR from the finger sensor. HR_rPPG_filtered is HR_rPPG after Kalman filtering

Figure 5. RPPG using cICA

benefit with motion compensation which itself is another subject for research. Finally, we average the entire image (or cropped or skin segmented version of it) to obtain a single value and thus loose any spatial information. Higher order analysis which preserves the spatial relationships between pixel neighborhoods can be an important avenue to look into.

References

[1] D. Bertsekas. Constrained optimization and Lagrange multiplier methods, 1982. 4

[2] P. Comon. Independent component analysis, A new concept? *Signal Processing*, 36(3):287–314, 1994. 1

[3] C. Ó. Conaire, N. E. O'Connor, and A. F. Smeaton. Detector adaptation by maximising agreement between independent data sources. *CVPR*, 2007. 5

[4] G. De Haan and V. Jeanne. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 1, 2

[5] G. de Haan and a. van Leest. Improved motion robustness of remote-PPG by using the blood volume pulse signature. *Physiological measurement*, 35(9):1913–1926, 2014. 2

[6] D. Djuwari, D. Kant Kumar, and M. Palaniswami. Limitations of ICA for Artefact Removal. *IEEE Medicine and Biology Society*, 5:4685–4688, 2005. 1

[7] A. B. Hertzman. Photoelectric Plethysmography of the Fingers and Toes in Man. *Experimental Biology and Medicine*, 37(3):529–534, 1937. 1

[8] A. Hyvärinen and E. Oja. A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation*, 9(7):1483–1492, 1997. 3

[9] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000. 1, 3

[10] P. a. Karjalainen. An advanced detrending method with application to HRV analysis Mika P. Tarvainen, Perttu O. Ranta-aho, and Pasi A. Karjalainen. pages 1–4. 2

[11] W. Lu and J. C. Rajapakse. ICA with Reference. 2, 4

[12] W. Lu and J. C. Rajapakse. Rajapakse, âĂIJConstrained independent component analysis. 10:570–576, 2000. 4

[13] W. Lu and J. C. Rajapakse. Approach and applications of constrained ICA. *IEEE Transactions on Neural Networks*, 16(1):203–212, 2005. 1

[14] D. McDuff, S. Gontarek, and R. W. Picard. Improvements in Remote Cardio-Pulmonary Measurement Using a Five Band Digital Camera. *IEEE transactions on bio-medical engineering*, 9294(10):1–8, 2014. 1, 2

[15] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Noncontact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 1, 2

[16] M. Z. Poh, D. J. McDuff, and R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, 2011. 1

[17] J. C. Rajapakse, F. Kruggel, J. M. Maisog, and D. Y. von Cramon. Modeling hemodynamic response for analysis of functional MRI time-series. *Human brain mapping*, 6(4):283–300, 1998. 1

[18] W. Verkruysse, L. O. Svaasand, and J. S. Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 1

[19] W. Wang, S. Stuijk, and G. D. Haan. Exploiting Spatial Redundancy of Image Sensor for Motion Robust rPPG. 62(2):415–425, 2015. 2

Appendix A. Derivatives of $g(w)$

Here we present the first and second derivatives of $g(w)$ needed by the lagrange multipliers method. We follow the convention that the derivative of a scalar w.r.t a column vector is a column vector of the same size as that of the vector. The first derivative of $g(w)$ in equation 9 can be obtained as follows considering squared autocorrelation as $\mathbf{r}^2 = [r_1^2\, r_2^2 \ldots r_N^2]$.

$$g'(\mathbf{w}) = -E\{\frac{\partial}{\partial \mathbf{w}}(\begin{bmatrix} r_1^2 & r_2^2 & \cdots & r_N^2 \end{bmatrix})\} \tag{18}$$

where the derivative of the squared autocorrelation $\mathbf{r}^2$ is then obtained using the chain rule of derivatives. Also, we know that $\mathbf{y} = \mathbf{w}^T\mathbf{x}$ giving $\frac{\partial \mathbf{y}}{\partial \mathbf{w}} = \mathbf{x}$.

$$\frac{\partial(\mathbf{r}^2)}{\partial \mathbf{w}} = \mathbf{x}\frac{\partial(\mathbf{r}^2)}{\partial \mathbf{y}} \tag{19}$$

$$= \mathbf{x}\begin{bmatrix} \frac{\partial(r_1^2)}{\partial y_1} & \frac{\partial(r_2^2)}{\partial y_1} & \cdots & \frac{\partial(r_N^2)}{\partial y_1} \\ \frac{\partial(r_1^2)}{\partial y_2} & \frac{\partial(r_2^2)}{\partial y_2} & \cdots & \frac{\partial(r_N^2)}{\partial y_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial(r_1^2)}{\partial y_N} & \frac{\partial(r_2^2)}{\partial y_N} & \cdots & \frac{\partial(r_N^2)}{\partial y_N} \end{bmatrix} \tag{20}$$

$$= \mathbf{x}\begin{bmatrix} 2r_1\frac{\partial r_1}{\partial y_1} & \cdots & 2r_N\frac{\partial r_N}{\partial y_1} \\ \vdots & \ddots & \vdots \\ 2r_1\frac{\partial r_1}{\partial y_N} & \cdots & 2r_N\frac{\partial r_N}{\partial y_N} \end{bmatrix} \tag{21}$$

$$= 2\mathbf{x}\begin{bmatrix} r_1\frac{\partial r_1}{\partial y_1} & \cdots & r_N\frac{\partial r_N}{\partial y_1} \\ \vdots & \ddots & \vdots \\ r_1\frac{\partial r_1}{\partial y_N} & \cdots & r_N\frac{\partial r_N}{\partial y_N} \end{bmatrix} \tag{22}$$

The size of $\frac{\partial(\mathbf{r}^2)}{\partial \mathbf{w}}$ is then $3 \times N$ from the product of $\mathbf{x}_{3\times N}$ with the jacobian of size $N \times N$. Consequently, its expectation ends up having a size of 3×1 since it is nothing but a temporal mean over N samples. The jacobian in equation 22 can be concisely expressed as $\begin{bmatrix} r_1\frac{\partial r_1}{\partial \mathbf{y}} & r_2\frac{\partial r_2}{\partial \mathbf{y}} & \cdots & r_N\frac{\partial r_N}{\partial \mathbf{y}} \end{bmatrix}$ where each column is the product of the derivative $\frac{\partial r_1}{\partial \mathbf{y}}$ and the scalar r_k and is of size $N \times 1$. Deriving equation 4, $r_k = \mathbf{y}T_k\mathbf{y}^T$, listed here for convenience, w.r.t $\mathbf{y}$ using the product rule of differentiation,[1]

$$\begin{aligned} \frac{\partial r_k}{\partial \mathbf{y}} &= \mathbf{y}\frac{\partial}{\partial \mathbf{y}}(T_k\mathbf{y}^T) + \mathbf{y}T_k\frac{\partial}{\partial \mathbf{y}}(\mathbf{y}^T) \\ &= \mathbf{y}\frac{\partial}{\partial \mathbf{y}}(\mathbf{y}T_k^T) + \mathbf{y}T_k \\ &= \mathbf{y}T_k^T + \mathbf{y}T_k = \mathbf{y}(T_k^T + T_k) \end{aligned} \tag{23}$$

where $\frac{\partial}{\partial \mathbf{y}}(T_k\mathbf{y}^T) = T_k^T$ comes from the fact that the differential of $T_k\mathbf{y}^T$, a vector, will remain the same even when it is transposed and the derivative is computed element-wise. For conciseness, we will represent the sum $T_k + T_k^T$ as $\mathbf{T}_k$. Finally to be consistent with our convention, using the same argument of the differential being immutable to transpositions, the row vector $\frac{\partial r_k}{\partial \mathbf{y}}$ can be transposed into a column vector and the matrix $\frac{\partial \mathbf{r}}{\partial \mathbf{y}}$ can be built as

$$\frac{\partial \mathbf{r}}{\partial \mathbf{y}} = \begin{bmatrix} r_1\mathbf{T}_1\mathbf{y}^T & \cdots & r_N\mathbf{T}_N\mathbf{y}^T \end{bmatrix} \tag{24}$$

[1] This result is owing to the fact that T_k is not symmetric. If it were symmetric, then the result would have been $2\mathbf{y}T_k$.

giving $g'(\mathbf{w})$ in equation 18 as

$$g'(\mathbf{w}) = -2\mathbf{x}E\left\{\begin{bmatrix} r_1\mathbf{T}_1\mathbf{y}^T & \cdots & r_N\mathbf{T}_N\mathbf{y}^T \end{bmatrix}\right\}$$

which can be further simplified to

$$g'(\mathbf{w}) = -2\mathbf{x}\begin{bmatrix} \mathbf{T}_1\mathbf{y}^T & \cdots & \mathbf{T}_N\mathbf{y}^T \end{bmatrix}\mathbf{r}^T/N \tag{25}$$

Since the expectation is a temporal mean the element-wise multiplication with r_k can be replaced by multiplication with the vector $\mathbf{r}^T$ which also simplifies the computation.

Next, to simplify the calculation of the second derivative of $g(\mathbf{w})$, we perform column-wise matrix multiplication in equation 25, omitting the scalar multiplication and division, to obtain

$$g'(\mathbf{w}) = -\mathbf{x}\begin{bmatrix} \mathbf{T}_1 r_1\mathbf{y}^T + & \cdots & +\mathbf{T}_N r_N\mathbf{y}^T \end{bmatrix} \tag{26}$$

$$= -\mathbf{x}\sum_{k=0}^{N}\mathbf{T}_k r_k\mathbf{y}^T \tag{27}$$

And since differentiation and summation are interchangeable based on the sum rule, $g''(w)$ can be obtained by

$$g''(\mathbf{w}) = -\mathbf{x}\sum_{k=0}^{N}\frac{\partial(\mathbf{T}_k r_k\mathbf{y}^T)}{\partial \mathbf{w}} \tag{28}$$

$$= -\mathbf{x}\sum_{k=0}^{N}\frac{\partial(\mathbf{T}_k r_k\mathbf{y}^T)}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{w}} \tag{29}$$

$$= -\mathbf{x}\left(\sum_{k=0}^{\mathbf{N}}\frac{\partial(\mathbf{T_k r_k y^T})}{\partial \mathbf{y}}\right)\mathbf{x}^T \tag{30}$$

The derivative of $\mathbf{T}_k r_k\mathbf{y}^T$ w.r.t $\mathbf{y}$ is then obtained by the product rule of differentiation.

$$\frac{\partial(\mathbf{T}_k r_k\mathbf{y}^T)}{\partial \mathbf{y}} = \mathbf{T}_k\frac{\partial r_k}{\partial \mathbf{y}}\mathbf{y}^T + \mathbf{T}_k r_k \tag{31}$$

$$= \mathbf{T}_k\left(\frac{\partial r_k}{\partial \mathbf{y}}\mathbf{y}^T + r_k\right) \tag{32}$$

which is of size $N \times N$. Consequently, the size of $g''(\mathbf{w})$ turns out to be 3×3 since the sum of $\frac{\partial(\mathbf{T}_k r_k\mathbf{y}^T)}{\partial \mathbf{y}}$ over N samples is also of size $N \times N$. $g''(\mathbf{w})$ is further used in the lagrange multipliers method for cICA , implemented as an augmented lagrangian method,presented in the next section.

Remote Photoplethysmography Based on Implicit Living Skin Tissue Segmentation

Serge Bobbia, Yannick Benezeth, Julien Dubois
Univ. Bourgogne Franche-Comté
LE2I UMR6306, CNRS, ENSAM
F-21000 Dijon, France
Email: serge.bobbia@u-bourgogne.fr

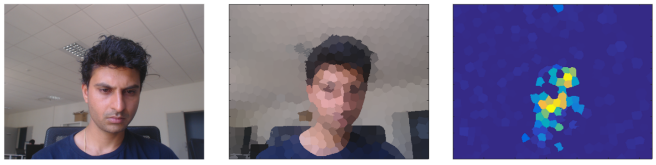
***Abstract*—Region of interest selection is an essential part for remote photoplethysmography (rPPG) algorithms. Most of the time, face detection provided by a supervised learning of physical appearance features coupled with skin detection is used for region of interest selection. However, both methods have several limitations and we propose to implicitly select living skin tissue via their particular pulsatility feature. The input video stream is decomposed into several temporal superpixels from which pulse signals are extracted. Pulsatility measure for each temporal superpixel is then used to merge pulse traces and estimate the photoplethysmogram signal. This allows to select skin tissue and furthermore to favor areas where the pulse trace is more predominant. Experimental results showed that our method perform better than state of the art algorithms without any critical face or skin detection.**

I. Introduction

Photoplethysmography (PPG) is a non-invasive technique for detecting microvascular blood volume changes in tissues. Nowadays, PPG is applied ubiquitously in many settings where a contact PPG sensor (also known as pulse oximeter) is typically attached to a finger or patched to the skin. Basically, contact PPG sensors are used to determine the heart rate and oxygen saturation in blood. The principle of this technology is actually very simple as it only requires a light source and a photodetector. The light source illuminates the tissue and the photodetector measures the small variations in transmitted or reflected light associated with changes in perfusion in the tissue [1].

However, conventional contact PPG sensor is not suitable in situations of skin damage or when unconstrained movement is required. Moreover, it has been showed that pressure of the conventional clip sensors tend to affect the waveform of PPG signal because of the contact force between the finger and the sensor [2].

With the emergence of camera-based health care monitoring, remote photoplethysmography (rPPG) has recently been developed as it allows remote physiological measurement without expensive and specialist hardware. Actually, it has been shown recently [3] that it is possible to recover the cardiovascular pulse wave measuring variations of back-scattered light remotely, using only ambient light and low-cost vision systems. Since this seminal work, there has been rapid growth in the literature pertaining to remote PPG techniques.



(a) (b) (c)

Fig. 1. Pulsatility measures estimated from various temporal superpixels. (a) input frame, (b) temporal superpixel segmentation and (c) pulsatility measures (blue means low pulsatility measures and yellow/orange is high).

Most methods share a common pipeline-based framework (*e.g.* [4]–[7]): regions of interest (ROI) are first detected and tracked over frames, RGB channels are then combined to estimate the pulse signal, which is filtered and analyzed to extract physiological parameters such as heart rate or respiration rate. An interesting and comprehensive state of the art paper on PPG and rPPG has been recently proposed by Sun and Thakor [1]. The selection of ROI is a critical first step to obtain reliable pulse signals. The ROI should contain as many skin pixels as possible. Several approaches have been proposed for ROI selection in the video stream. In earlier studies, manual selection of the ROI have been used [3], [8]. ROI can also be defined based on the result of classical face detection [9] and tracking [10] algorithms and possibly refined with skin pixel classification [11]. It is worth mentioning that this preliminary step can be computationally very expensive (*e.g.* [11]).

Pixels in the ROI are then usually spatially averaged and the process is repeated in each video frame. The result of this process is a time series, that is later used to obtain rPPG signal. It has been shown in several studies that the quality of the ROI has a direct impact on the quality of the rPPG signal (*e.g.* in [12]). First, because a smaller number of skin pixels leads to larger quantized RGB errors, it can be observed that the quality of rPPG signal deteriorates while down-sampling the ROI. This may be understood as the reduction of the sensor noise amplitude by a factor equal to the square root of the number of pixels used in the averaging process [13]. Second, the quality is also affected by the percentage of non-skin pixels in the ROI [7]. All rPPG algorithms suffer from performance degradation when the ROI is not properly selected. These two remarks are fairly intuitive but it is actually quite difficult in

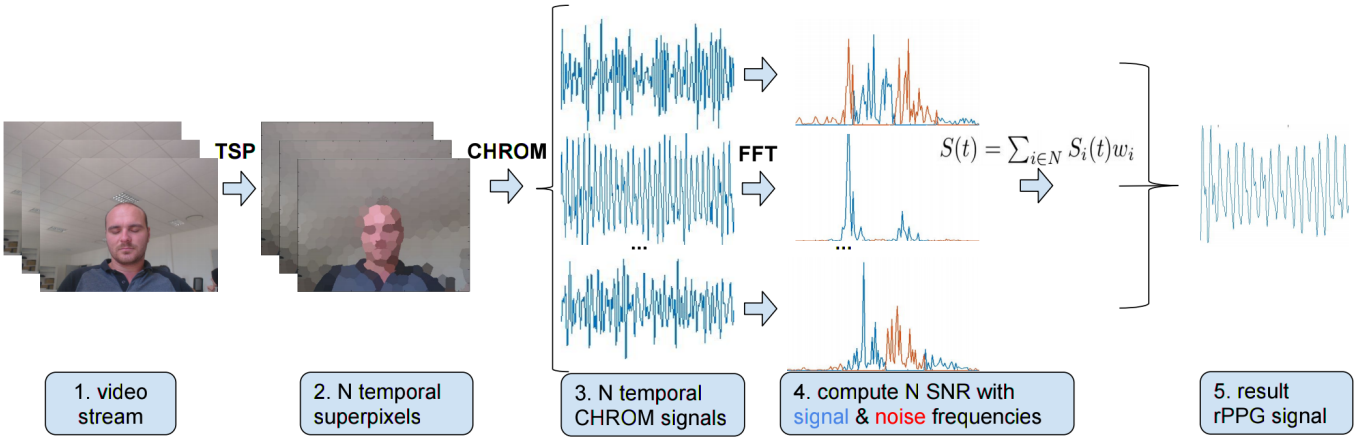


Fig. 2. Overview of the proposed method. (1) Input video stream is (2) decomposed into temporally consistent superpixels. (3) *Tentative* rPPG signal is extracted from each TSP. (4) A pulsatility measure is estimated for each ROI. Blue signal is the convolution of the periodogram by h_{signal} and red signal is the convolution by h_{noise}. (5) A weighted average of all the *tentative* rPPG signals is finally computed.

practice to get a well-defined ROI, that is stable over time, without performing complex calculations. Moreover, as shown by [14], the rPPG signal is not distributed homogeneously on skin. Some skin regions contain more PPG signal than others. For example, we experimentally observed that signal-to-noise ratio (SNR) of photoplethysmogram signals extracted from forehead or cheekbones are clearly higher than those obtained from the chin. Figure 1 presents the SNR of rPPG signals calculated from several skin regions.

From these observations, we propose a new method that implicitly selects ROI that represents living skin tissue and that favor regions of interest where the pulse trace is more predominant. We use the term *implicit* to differentiate our method with those that require critical pre-processing steps for ROI selection and tracking. As opposed to conventional approaches based on face detection, tracking and skin segmentation, ROI selection is based on the fact that only the skin tissue of an alive subject exhibits pulsatility. The input video stream is decomposed into several temporal superpixels from which pulse signals are extracted. Pulsatility measure for each temporal superpixel is then used to merge pulse traces and estimate the photoplethysmogram signal. This approach can be associated with any rPPG algorithm. In this paper, we experimentally validated the proposed automatic living skin tissue segmentation for ROI selection using the chrominance-based method (also known as CHROM) [6] because this method is definitely one of the most reliable rPPG methods. To the best of our knowledge, work by [15], called Voxel-Pulse-Spectral (VPS) is the closest contribution to ours. They propose an automatic and unsupervised subject detection via rPPG. However computational load of VPS is significantly heavier. Moreover, our method uses temporal superpixels tailored to video data rather than supervoxels such as in VPS that are designed for 3D volumetric data. In contrast to supervoxel methods, with temporal superpixels object parts in different frames are tracked by the same temporal superpixel. Guazzi *et al.* [13] also use the fusion of several pulse traces but the video is simply divided into contiguous square blocks. The temporal superpixel segmentation is more suited to rPPG algorithms to handle motion scenarios.

The rest of the paper is organised as follows. The method is described in section II with the temporal superpixel segmentation, the pulsatility measure and the fusion procedure. The video dataset, metrics and results are described in section III while the conclusion is presented in section IV.

II. Method

The overview of the proposed method is shown in figure 2. The algorithm can be decomposed in three main steps: (1) video stream is first decomposed into temporally consistent superpixels (later called TSP for Temporal SuperPixels). Then, tentative rPPG signal is extracted from each TSP. (2) A pulsatility measure is estimated for each TSP and (3) a weighted average of all the rPPG signals is computed where weights are given by the pulsatility measure.

A. Temporal superpixel-based pulse extraction

The first step of our method is the segmentation of the video stream into temporally consistent superpixels. If a superpixel is a set of pixels that are local, coherent, and which preserve most of the structure necessary for segmentation [16], temporal superpixels can be defined as a set of video pixels that are local in space and track the same part of an object across time [17]. In this work, we use the TSP method proposed by Chang *et al.* [17] which we found to be a good compromise between precision and speed. This method is based on the Simple Linear Iterative Clustering (called SLIC) [18] decomposition. It has been shown that SLIC is among the fastest superpixel methods and is very efficient [19].

In order to extract tentative rPPG signal from each TSP, for each video frame, pixels in each TSP are spatially averaged. The result of this process is a set of N RGB time series $x_i^c(t)$,

where $c \in \{R, G, B\}$ is the color channel, t is the frame index and $i = 1, 2, \ldots, N$ with N is the number of TSP:

$$x_i^c(t) = \frac{\sum_{k=1}^{M_i(t)} I_{k,i}^c(t)}{M_i(t)} \tag{1}$$

where $M_i(t)$ is the number of pixels in the i^{th} TSP at time t and $I_{k,i}^c(t)$ the k^{th} pixel value at time t and color channel c.

The RGB temporal traces are then pre-processed by zero-mean and unit variance normalization, detrended using smoothness priors approach [20] and band-pass filtered with Butterworth filter. The rPPG signal is then extracted using the chrominance-based method (later called CHROM) [6]. This method applies simple linear combinations of RGB channels and obtains very interesting performance with low computational complexity. Let $y_i^c(t)$ be the RGB time series obtained after pre-processing, CHROM method projects RGB values onto two orthogonal chrominance vectors X_i and Y_i:

$$\begin{aligned} X_i(t) &= 3y_i^R(t) - 2y_i^G(t), \\ Y_i(t) &= 1.5y_i^R(t) + y_i^G(t) - 1.5y_i^B(t). \end{aligned} \tag{2}$$

The pulse signal S_i of the i^{th} TSP is finally calculated with $S_i(t) = X_i(t) - \alpha Y_i(t)$ where $\alpha_i = \sigma(X_i)/\sigma(Y_i)$. Because X_i and Y_i are two orthogonal chrominance signals, PPG-induced variations will likely be different in X_i and Y_i, while motion affects both chrominance signals identically. S_i is called a *tentative* rPPG signal because some pulse signals calculated from background superpixels do not contain relevant information.

B. Pulsatility measure

Only the skin tissue of an alive subject exhibits pulsatility, therefore pulse signals calculated from some superpixels only contain noise (*e.g.* on non-skin areas). Figure 3 (a) presents a periodogram of a pulse signal estimated from skin area while figure 3 (b) presents periodogram of a pulse signal estimated from the background. In the frequency domain, the pulsatile, cardiac-synchronous signal, exhibits an important peak centered on the fundamental frequency of heart rate, possibly its second harmonic and limited information at other frequencies. To measure the quality of rPPG signals, we estimate signal-to-noise ratio (SNR) defined as the ratio of the power of the main pulsatile component and the power of background noise, compute in dB due to the wide dynamic range of the signals.

The pulsatility measure of the i^{th} TSP is estimated by:

$$SNR_i = 10 \log_{10} \left(\frac{\int_{f_1}^{f_2} h_{signal}^i(f) |\mathcal{F}\{S_i(t)\}|^2 \mathrm{d}f}{\int_{f_1}^{f_2} h_{noise}^i(f) |\mathcal{F}\{S_i(t)\}|^2 \mathrm{d}f} \right) \tag{3}$$

where $\mathcal{F}\{S_i(t)\}$ is the Fourier transform of the rPPG signal of the i^{th} TSP, f_1 and f_2 the lower and upper limit of the integral defined by the possible physiological range of the heart rate

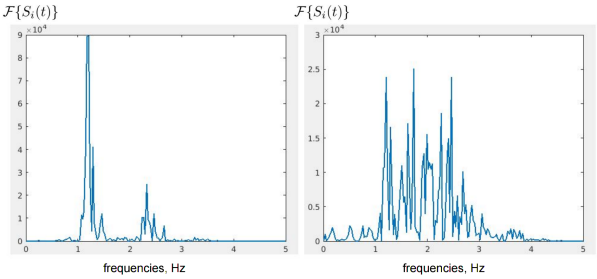


Fig. 3. Periodogram examples of 2 *tentative* rPPG signals estimated from (a) skin area and (b) background.

(40 to 240 bpm in our case), and a double-step function h, for the first and second harmonics, defined by the convolution:

$$\begin{aligned} h_{signal}^i(f) &= [\delta(f - f_0^i) + \delta(f - 2f_0^i)] * \prod(\pm f_r) \\ h_{noise}^i(f) &= 1 - h_{signal}^i(f) \end{aligned} \tag{4}$$

with δ the Dirac delta function, f_0^i the fundamental frequency (*i.e.* peak of the periodogram), convoluted with the $rect$ function, noted as $\prod$ of half-width f_r. SNR_i will be high for skin TSP and low for background ones.

C. rPPG signal fusion

The final rPPG signal $S(t)$ is obtained by a weighted average of all *tentative* pulse signals $S_i(t)$, *i.e.* $S(t) = \sum_{i \in N} S_i(t) w_i$ where weightings w_i are a function of the main pulsatile component SNR:

$$w_i = \frac{10^{SNR_i}}{\sum_{i \in N} 10^{SNR_i}} \tag{5}$$

Weights are normalized and in order to conserve the relative contribution of each rPPG signal, weights are defined with the $\log^{-1}(x)$ function (*i.e.* 10^x) The weighting favors TSP that have a high main pulsatile component SNR as these are more likely to represent skin areas. For example, in figure 4, the final rPPG signal is made up of mainly three *tentative* rPPG signals.

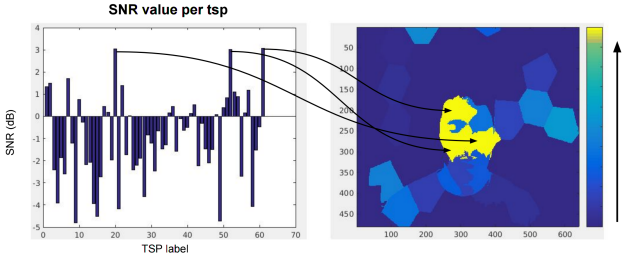


Fig. 4. Examples of SNR values (dB) and its corresponding superpixel.

III. Experiments

This section presents the experimental setup for evaluating the proposed method. First, we describe the video dataset. Then, we present the evaluation metrics and finally we compare our implicit ROI selection with regular face detection/tracking and skin detection approach.

A. Video dataset

Videos have been recorded using a simple low cost webcam (Logitech C920 HD Pro) at 30fps with a resolution of 640x480 in uncompressed 8-bit RGB format. During the recording, the subject sits in front of the camera (about 1m away from the camera) with his/her face visible. Participants were asked to sit still but some videos present significant movement (especially at the beginning of the sequence). The dataset is composed of 7 videos (about 16500 frames). All experiments are conducted indoors with a varying amount of sunlight and indoor illumination. To validate the heart rate estimations, PPG is recorded using transmissive pulse-oximeter finger clip sensor (Contec Medical CMS50E) and synchronized with the video. Video frames synchronized with PPG sensor data can be downloaded from our project page[1].

B. Benchmark algorithms and metrics

We compare heart rate estimation of our implicit ROI selection method with classical skin segmentation. To make the comparison with our method fair, chrominance-based method [6] is also used with the skin segmentation method (indistinctly called CHROM or reference method). For the reference method, skin has been segmented based on color feature using the implementation of Conaire *et al.* [21] after face detection [22] and tracking [23]. Examples of skin segmentation results are presented in figure 5. We use the same pre-processing and filtering for both methods.

Fig. 5. Skin segmentation result examples for the reference method.

For each video, we estimate heart rate in a sliding window framework. Welch's method is used to obtain the periodogram over a 30 second moving window, with a step size of one second. Heart rate is given by the position of the peaks on the frequency axis. The same heart rate estimation procedure was used on the PPG signal recorded with the contact sensor, on the rPPG signal given by the reference method and the rPPG signal given by our method.

The following metrics are used for comparison:

- **Bland-Altman plots** that measure the agreement between heart rate estimated from rPPG signals and PPG signal. The lines represent the mean and 95% limits of agreement.
- **Correlation plots** and the Pearson correlation value r^2 between heart rate estimated from rPPG signals and PPG signal.
- **Root mean square error** (RMSE).
- **Precision at 2.5 or 5 bpm**. This metric represents the percentage of estimations where the absolute error is under a threshold (2.5 or 5 bpm).

[1]http://ilt.u-bourgogne.fr/benezeth/projects/ICPR2016/

C. Results

In the first experiment, we evaluate the robustness of the proposed method to TSP resolution varying the number of temporal superpixels N per frame. N varies from 50 to 1000 (gross to fine segmentation). Figure 6 presents precision at 2.5 and 5 bpm. Results are stable, from 95% to 99% at 5bpm and from 89% to 95% at 2.5 bpm. Smaller number of pixels in a TSP leads to larger quantized RGB errors that seams to be compensated by the weighted average procedure. For the following experiments, we use $N = 200$.

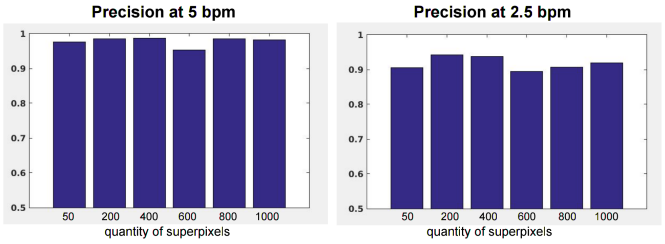


Fig. 6. Precision at 5 and 2.5 bpm calculated on all videos in percentage.

In the second experiment, we compare heart rate estimations with our method and the reference method (called CHROM). Figures 7 and 8 present correlation plots and Bland-Altman plots of the 2 methods. With 200 TSPs per frame, we obtain a Pearson correlation value of 0.9847 compared to 0.9793 for CHROM. From Bland-Altman plots, we can see that the mean bias is -0.36 bpm with 95% limits of agreement 3.4 to 2.7 bpm for our method and the mean bias is -0.55 bpm with 95% limits of agreement 4.1 to 3 bpm for CHROM. We can observe that results are actually very good, even for the reference method. Our method performs slightly better than CHROM. However, the small difference can be explained by the fact that results obtained with the reference are indeed "too good". In table I, we present a summary of all obtained results with a TSP resolution N of 200 and 400 comparedwith the reference. For all metrics, best results are always obtained by our method alternatively with $N = 200$ or $N = 400$.

TABLE I
Summary of results

Evaluation metrics	N=200	N=400	CHROM
Precision at 5 BPM	0.9856	**0.9869**	0.9786
Precision at 2.5 BPM	**0.9428**	0.9378	0.9386
RMSE	1.5	**1.43**	2.43
Pearson correlation r^2	**0.9847**	0.9843	0.9793

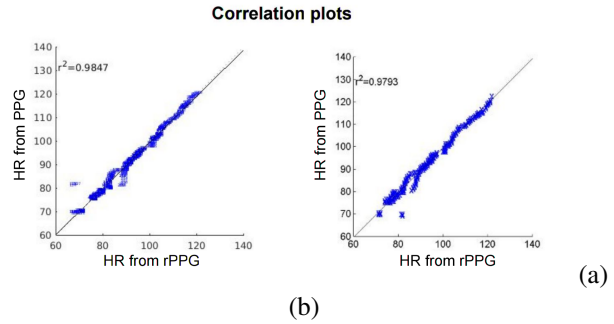


Fig. 7. Correlation plots obtained from all videos with (a) our method and (b) the baseline CHROM. Abscissa represents heart rate (HR) estimated from rPPG signal and ordinate represents HR estimated from the PPG signal (ground truth).

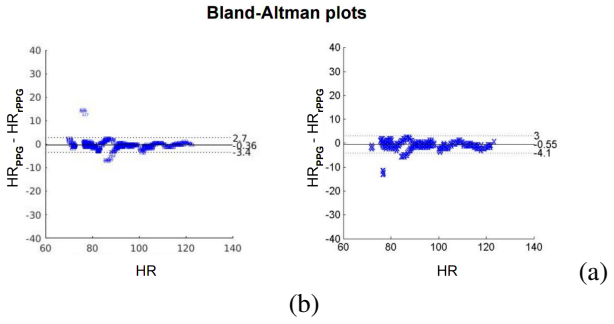


Fig. 8. Bland-Altman plot from all videos with (a) our method and (b) the baseline CHROM. Abscissas represent HR and ordinate represents the difference between HR estimated from PPG (HR_{PPG}) and HR estimated from rPPG (HR_{rPPG}).

IV. Conclusion

Remote photoplethysmography technology has a great potential for use in a wide range of clinical assessments, such as homecare, telemedicine and personal healthcare. The selection of ROI is a critical first step of rPPG techniques to obtain reliable pulse signals. In the present study, we have described, implemented, and evaluated a new rPPG method that implicitly select living skin tissue via their particular pulsatility feature. Photoplethysmogram signals are estimated with the weighted fusion of several *tentative* rPPG signals computed on a set of temporal superpixels. Even if the dataset is very simple, the results of this study have demonstrated that the rPPG signals could be remotely estimated without any tedious ROI selection. Our method always performs slightly better than the reference method CHROM.

Further developments include the evaluation of the proposed technique on a larger dataset with more challenging scenarios (motion, illumination variation, compression noise etc.). Also, we will continue the developments to handle cases where multiple persons are in the scene and we will evaluate our implicit ROI selection method with other rPPG algorithms.

Acknowledgment

This research was supported by the Conseil Régional de Bourgogne Franche-Comté, France and the Fond Européen de Développement Régional (FEDER).

References

[1] Y. Sun and N. Thakor, Photoplethysmography Revisited: From Contact to Noncontact, From Point to Imaging, IEEE Trans. on Biomedical Engineering 63 (3) (2016) 463 - 477.

[2] X.F. Teng and Y.T. Zhang, The effect of contacting force on photoplethysmographic signals, Physiological Measurement 25 (5) (2004) 1323-1335.

[3] W. Verkruysse, L. O. Svaasand, J. S. Nelson, Remote plethysmographic imaging using ambient light, Optics express 16 (26) (2008) 21434-21445.

[4] M.Z. Poh, D.J. McDuff and R.W. Picard, Non-contact automated cardiac pulse measurements using video imaging and blind source separation, Optics Express 18 (10) (2010) 10762-10774.

[5] B.S. Kim and S.K. Yoo, Motion artifact reduction in photoplethysmography using independent component analysis, IEEE Trans. on Biomedical Engineering 53 (3) (2006) 566-568.

[6] G. de Haan and V. Jeanne, Robust pulse rate from chrominance-based rppg, IEEE Trans. on Biomedical Engineering 60 (10) (2013) 2878-2886.

[7] W. Wang, S. Stuijk and G. de Haan, A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation, IEEE Trans. on Biomedical Engineering (2015).

[8] Y. Sun, S. Hu, V. Azorin-Peris, S. Greenwald, J. Chambers and Y. Zhu, Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise, Journal of Biomedical Optics 16 (7) (2011).

[9] M. Z. Poh, D. J. McDuff and R. W. Picard, Advancements in non- contact, multiparameter physiological measurements using a webcam. IEEE Trans. on Biomedical Engineering 58 (1) 2011 7-11.

[10] H.E. Tasli, A. Gudi and M. Uyl, Remote PPG based vital sign measurement using adaptive facial regions, In IEEE International Conference on Image Processing (2014) 1410-1414.

[11] W. Wang, S. Stuijk and G. de Haan, Exploiting Spatial-redundancy of Image Sensor for Motion Robust rPPG, IEEE Trans. On Biomedical Engineering 62 (2) (2015) 415-425.

[12] F. Bousefsaf, C. Maaoui, A. Pruski, Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate, Biomedical Signal Processing and Control 8 (6) (2013) 568-574.

[13] A.R. Guazzi, M. Villarroel, J. Jorge, J. Daly, M.C. Frise, P.A. Robbins and L. Tarassenko, Non-contact measurement of oxygen saturation with an RGB camera, Biomedical Optics Express 6 (9) (2015) 3320-3338.

[14] A.A. Kamshilin1, E. Nippolainen, I.S. Sidorov, P.V. Vasilev, N.P. Erofeev, N.P. Podolian and R.V. Romashko, A new look at the essence of the imaging photoplethysmography Scientific Reports 5 (2015).

[15] W. Wang, S. Stuijk, and G. de Haan, Unsupervised Subject Detection via Remote PPG, IEEE Trans. On Biomedical Engineering 62 (11) (2015) 2629-2637.

[16] X. Ren and J. Malik, Learning a classification model for segmentation, In Proc. IEEE Conference on Computer Vision and Pattern Recognition 1 (2003) 10-17.

[17] J. Chang, D. Wei and J. Fisher, A video representation using temporal superpixels, In Proc. IEEE Conference on Computer Vision and Pattern Recognition (2013) 2051-2058.

[18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk, SLIC Superpixels Compared to State-of-the-art Superpixel Methods, IEEE Trans. on Pattern Analysis and Machine Intelligence 34 (11) (2012) 2274-2282.

[19] P. Neubert and P. Protzel, Superpixel benchmark and comparison, In Proceedings Forum Bildverarbeitung, (2012) 1-12.

[20] M.P. Tarvainen, P.O. Ranta-Aho and P.A. Karjalainen, An advanced detrending method with application to HRV analysis, IEEE Trans. on Biomedical Engineering 49 (2) (2002) 172-175.

[21] C.O. Conaire, N. O'Connor and A.F. Smeaton, Detector adaptation by maximising agreement between independent data sources, IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum (2007).

[22] P. Viola and M. Jones, Rapid Object Detection using a Boosted Cascade of Simple Features, IEEE Conference on Computer Vision and Pattern Recognition 1 (2001) 511-518.

[23] C. Tomasi and T. Kanade, Detection and Tracking of Point Features, Carnegie Mellon University Technical Report CMU-CS-91-132 (1991).

COMPARISON BETWEEN GENETIC ALGORITHM AND PARTICLE SWARM OPTIMIZATION FOR POSITIONING A SET OF CAMERAS FOR VIDEO SURVEILLANCE

David Strubel [1;2] Mark Bastourous[1] Olivier Morel[1] Naufal M. Saad [2] David Fofi[1]

Abstract— **This paper have to aims of optimizing a vast coverage area to allow an image reconstruction using mosaicing techniques. Among the investigated methods to find the best camera positions, two of them are studied, namely the Particle Swarm Optimization (PSO) and the Genetic Algorithms (GA). After having performed many experiments to compare the algortithms, the GA is chosen for this performance and this adaptability to the problem. To validate the proposed method, it is applied on area of irregular shape and with the cameras mounted on the Unmanned Aerial Vehicles (UAVs). V-REP [14] is used to simulate the UAVs in the environment (indoor or outdoor). The simulation validates the efficiency of the proposed method to find the number and the poses of the cameras. Then by using the images acquired it is possible to compute mosaic images and control the area.**

Keyword: Coverage, Genetic Algorithm, Optimization, Mosaic, UAV, Application

I. INTRODUCTION

The optimal positioning of a cameras network is tricky and no efficient solution exists to solve it in a complex environment. The aim of this work is to provide a flexible and tunable solution for this problem and to analyse the performance of the current state-of-the-art techniques. The final objective of our research is to design a global optimization scheme allowing a camera network to self-organize and self-reconfigure, according to set of fixed priorities and constraints, in order to ensure a maximum coverage. Self-organization have to be performed to be efficient in realistic conditions. Within this context, it is important to assess the actual performance and limits of the state-of-the-art algorithms. First, we study the following contribution study and compare three standard algorithms, namely the Random Selection (RS), the Particle Swarm Optimization (PSO) and Genetic Algorithm (GA), in terms of efficiency and quality. The quality is evaluated with the coverage rate and the efficiency by the numbers of iteration for the optimisation to converge.
In addition the solution proposed is supported by an adapted and optimized cost function, specially adjust to the UAV problems.
We propose :

- A comparison between different algorithm from literature and the same family.
- Using the Genetic Algorithm for position many cameras in a big area.
- A cost function adapted to the problem of coverage using UAV.

1 LE2I UMR6306, CNRS, ENSAM, Univ. Bourgogne Franche-Comte
2 CISIR, Universiti Teknologi Petronas

II. RELATED WORKS

Sensor positioning problem has been investigated since in the last decades, mainly for video surveillance [1]. Without any additional constraint, this problem is a NP-Hard as stated in [1]–[5] Two solutions have been proposed to optimize the coverage of the area with a sensors network. The first one is based in the Art Galleries Problem (AGP) [1], [2]. AGP the position of body guard is optimized in a museum. The second is based on the Wireless Sensors Networks [6]–[9] trying to find the best position to design an efficient network which can collect data with any kind of sensors. However, the solution proposed to this problem works only with some constraints such as the sensor has 360 field of view and no obstacle. One of the algorithm commonly used is PSO as detailed in [10], [11]. In Zhou et al [10], some experimental results are provided and one solution running in real time is proposed. However, the scene used for the experiments is rather small and many cameras are employed to fully cover it. On the other hand, Reddy et al [11] optimize the position of the cameras by using a cost function but also handling resolution and lighting. The multi-goal approach affect the final solution. In Boeringer et al [11] also introduces the concept of acceptable response, allowing non-optimal/sub-optimal solutions. If the coverage score is good enough, the solution is accepted and not locked by the research of an optimal solution. Our paper is based on [10], [11], attempting to extend it by testing GA and PSO on different environment (basic room, big room, non-square shape) as well as for drone positioning.

III. OPTIMISATION OF CAMERAS POSITIONING

A. Objectives

The main purpose of our work is to estimate the position of n cameras surveying a given area in order to maximize the visual coverage. The cameras are mounted on a UAV and that each camera is assume to provide a top view of the area. Each cameras coverage is defined by the projection of the visual field onto the ground. Accordingly the combination of all the acquired images from the cameras allows to build a mosaic of the area to be checked. In order to assess the best

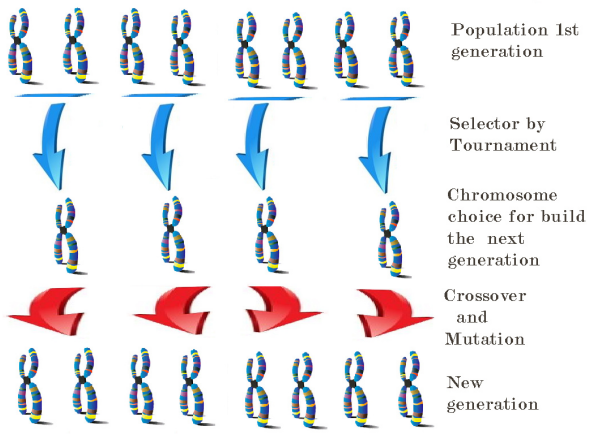


Fig. 1. GA explanation, from a generation to another.

algorithm among PSO and GA have been investigated. The following subsections will give an overview of both.

B. Particle Swarm Optimization

PSO is an evolutionary method [17], which aims at optimizing a problem by iteratively improving a candidate called particle. The quality of a particle is evaluated by a cost function. Thereby, the best particles are selected for the next generation. A new set of particles is defined by randomly moving around the best solution given by the previous iteration. The search-space around the best solution is controlled by a parameter called inertia. Repeating this method at every iteration push the algorithm to converge. To use properly the PSO few parameters need to be defined:

- The number of particles for each iterations.
- How the particles are initialize.
- The value of inertia.
- The cost function.

After several test on the different environments described in section IV.B the PSO are configured with an inertia of 0.5, the number of particle are 100, and the initialization of the paricles are fully random. Moreover the cost function used for the PSO is defined in section III.D

C. Genetic Algorithm

Motivated by Darwin′s theory of evolution and the concept of natural selection, the GA use processes analogous to genetic recombination and mutation. To promote the evolution of a population to reach a predefined goal [12], [18] . Such kind of algorithms require the definition of a genetic representation of the problem and the cost function is used to evaluate the solution. The candidate solution is represented by a data structure named chromosome, which is the equivalent of a particle for the PSO. The cost function and the data structure is the same for all the algorithms. Mostly it is to compare properly the algorithms.

The genetic algorithm work as iterative process. Every iteration are called generation and the chromosome of the actual generation are the offspering of the previous generation (see Fig.1. To pass from an generation to the other a few steps are necessary (see Figure 1):

1. For each sub-set, a selection is made, in order to select the most attractive chromosome (i.e. the one that maximizes the cost function).
2. Basic operations as cross-over and mutation are performed, on the selected chromosome to give rise to the next generation. [15]
3. The process is repeated until the convergence point or the generation boundary are reached. GAs allow more flexibility than PSOs however require more parameters to be configured.

In our experiments, we fixed the number of chromosomes to be 90, the operator used are the mutation and cross over with the mutation rate to be 0.001 and the crossover rate to be 0.919. The initial population are fully random and the selection is done by using a tournament selection method [16].

D. Cost function

Since the goal is to maximize the visual coverage of the camera network. They are expressed as a list of the cameras parameters as in the equation (1)

$$Vs = \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix}_{1 \leq i \leq n} \tag{1}$$

Where n is the number of cameras.
$x; y; z$ are the position of the camera in the Cartesian coordinate system.
The x and y are limited to the boundary fixed depending the area:

$$\begin{matrix} 0 \leq x \leq \text{width area} \\ 0 \leq y \leq \text{heigth area} \end{matrix} x, y \in \mathbb{N} \tag{2}$$

The z are chosen in a list of possible altitude predefined.

Thereafter the cost function is designed to qualify the solution, as follows:

$$\sum_{i=1}^{n} = \frac{cover(i)}{size(grid)}_{1 \leq i \leq n} \tag{3}$$

Where n is the number of cameras; $Grid$ represents the discretization of the ground plane (floor);
$Cover()$ is a function which computes the area on the ground which is covered by at least one camera;
$Size()$ is the dimension of the full area which must be covered.
Camera projection model is not explicitly taken into account, but the ground-projected visual field instead, as described in Figure 2. The equation (3) of the cost function can be also summarize in a basic pseudo-code in Algortihm : Coverage computation.

IV. Experimentation

A. Context of experimentation

In order to compare the two algorithms and evaluate their performances, many test on different scenarios depicted in

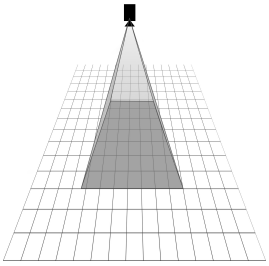


Fig. 2. Projection of the camera on the ground.

Algorithm . Coverage computation

```
procedure COST FUNCTION(Cameras)
    covered = 0
    while i = 1 ; i > size of the Grid do
        while j = 1 ; j > number of cameras do
            if Grid_i is visible by Camera_j then
                covered = Grid_i
    return ...
```

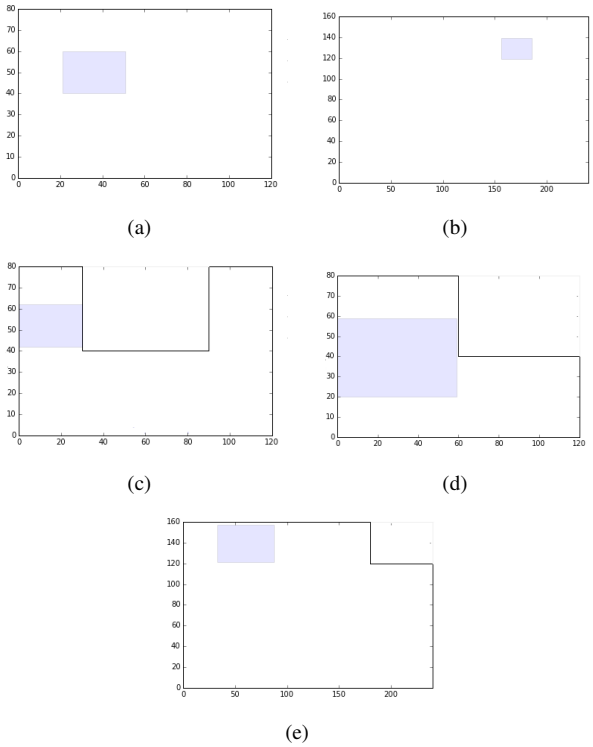
Figure 3, with have different sizes and shapes of rooms, where:

- z is the height of the camera between (within the range [1z ;z].
- Figure 3.a is an area of size 12080 (named Room).
- Figure 3.b is an area of size 240160 (named Big room)
- Figure 3.c is an area of size 12080 (named Room U)
- Figure 3.d is an area of size 12080 (named Room L)
- Figure 3.e is an area of size 24080 (named Big room L)

z=1		GA		PSO		RS	
		GT	NC	GT	NC	GT	NC
Room	120x80	16	20	16	20	16	20
	240x160	64	70	64	70	64	70
Room U	120x80	12	20	12	20	12	20
z=2		GA		PSO		RS	
Room	120x80	4	10	4	10	4	10
	240x160	16	20	16	20	16	20
Room L	120x80	3	10	3	10	3	10
	240x160	15	20	15	20	15	20

TABLE I
DESIGN OF EXPERIMENT FOR COMPARE THE EFFICIENCY OF RS, PSO AND GA IN DIFFERENT CONDITION. (GT IS GROUND TRUTH AND NC IS NUMBER OF CAMERAS).

The design of experiments (see table 1) is designed to identify the most efficient algorithm for the positioning of a set of cameras. GT that represents the Ground Truth is the minimum number of the cameras required to fully cover a given area. The size of the area has been selected so that the GT can be easily estimated. NC is the maximum number of cameras used for the experiments. For each experiment a solution is computed for a number of cameras from 1 to NC. Note that the PSO and GA are also compared with the Random Selection algorithm(RS), to ensure their performances are better than a pure random process. In order to compare the different algorithms in similar conditions, only 10000 calls of the cost function is allowed for each set of cameras.



(a) (b) (c) (d) (e)

Fig. 3. The scenarios used for the experiments: (a), (b), (c) are with z=1 and (d), (e) with z=2. The grey rectangle represents the field of view of one camera projected onto the ground.

B. Analyse of the results

After having performed the design of experiment (see table I) it appears the RS is always the worst solution (more obviously on Figure 4 Figure 6), as excepted we compare PSO and GA with RS to check their performance with supposedly the worst possible solution. The GA and PSO algorithm are close but the result was very different according to the parameters of the experiment. In some case the GA is more efficient (example in Figure 5) particularly in the case where the search space is large. Instead PSO is much more effective to optimize small areas (example in Figure 6). This efficiency is explained by the small variety of solution introduced by the PSO. However in the mean time, this small variety is perfect to optimize fastly in a small search space or when the local minima is not too much important. Although the variety introduced by the GA to pass over the local minima. This variety impacts the optimization accuracy and may have more difficulties to find an adjusted position. PSO is efficient when the area of the room is reduced whereas the GA is efficient for larger rooms. From the results obtained by both algorithms a simulation has been performed using

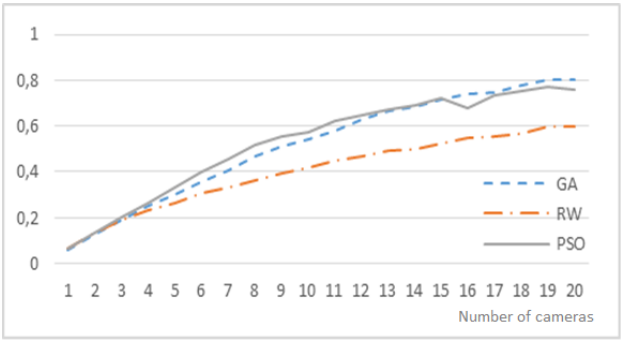


Fig. 4. Comparison of GA, RS, PSO algorithms with a Z between [1/2; 2], in the big room with L shape 240x160 and ground truth equal to 15.

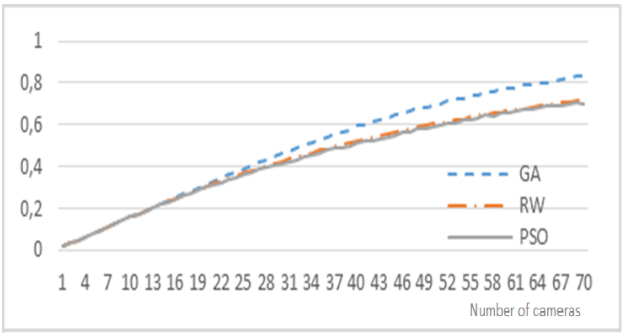


Fig. 5. Comparison of GA, RS, PSO algorithms with a Z equal to 1, in the big room 240x160 and ground truth equal to 64.

the robotics simulation tool VREP as presented Figure 7.

V. Application and other results

Thanks to the preliminaries studies the GA seems more suitable to optimize the coverage problem. Based on this finding other experiment was made in other sizes and complexity of environments. the optimisation and the coverage rate instead of complex area stay similar to the previous experiment which are did it with stochastic algorithm as GA and PSO. In addition these algorithm does not use heuristic, so their behavior stay the same and are not depending to the area to cover. The following experiment are done in mush bigger search space and a more complex area although the GA was used with the same parameter excepted the number of generation was free, the GA stop when the convergence point are reached.

A. Indoor coverage

The indoor experiment validate the efficiency of the GA for a same search space then the precedent section with a bit bigger amount of cameras although the area to cover is much more complex as in the Figure 8.

B. Outdoor coverage

The outdoor experiment helping to show the efficiency of GA in a bigger area and in a realistic environment. In the Figure 9.a the area to control form GPS images. From the GPS images the shape of the area to cover are extracted (see Figure 9.b) and the GA are performed to optimize the

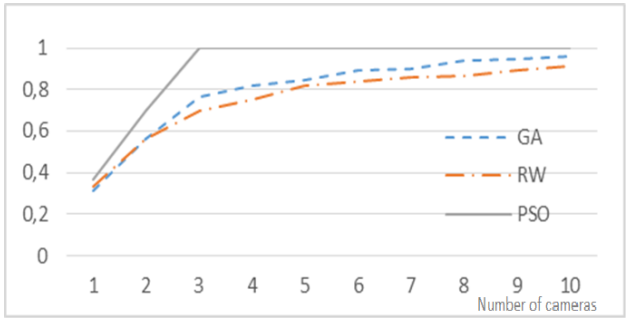


Fig. 6. Comparison of GA, RS, PSO algorithms with a Z between [1/2; 2], in the room with L shape 120x80 and ground truth equal to 15.

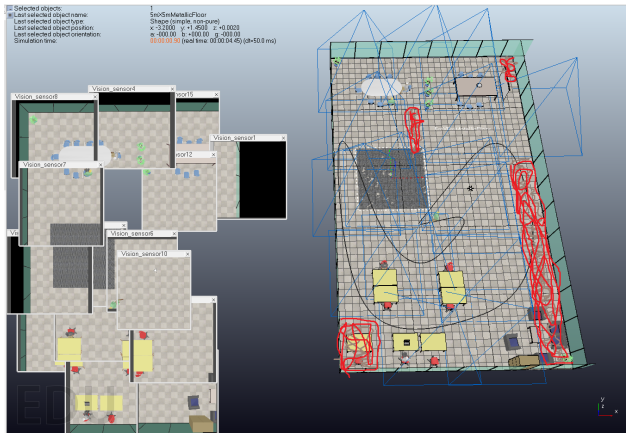


Fig. 7. Demonstration of the simulation with a robot simulation. This demonstration is with a set of 15 cameras and at the left side the mosaicking of the cameras is visible.

position of the cameras network the result are in visible in Figure 9.c.

VI. Conclusion

In this paper, the problem of coverage for video surveillance was formalized as an optimization problem. Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) were investigated and compared. Each algorithms has its own drawbacks and advantages:

PSO is more efficient in small environments. Indeed, when the inertia is relatively small, then from an iteration to the other, the algorithm evolves in a close neighborhood. While in contrast, when the inertia is bigger, the algorithm tends to behave like a random selection.

GA is more generic and more efficient, but its parameters have to be adapted to the specific configuration. The future work will be focus on continuing to optimize the coverage and to adapt the solution in the robotic context and application. The optimization of the coverage could take the direction to hybridate GA and PSO in order to combine the advantage of both algorithms or to continue only with the GA but with an adapted and with dynamic parameters. In the other direction the actual result may be extend to the problem of coverage path planing, with a Unmanned Aerial Vehicle to do the control of the area. Also this extension will be linked with the mosaicking application.

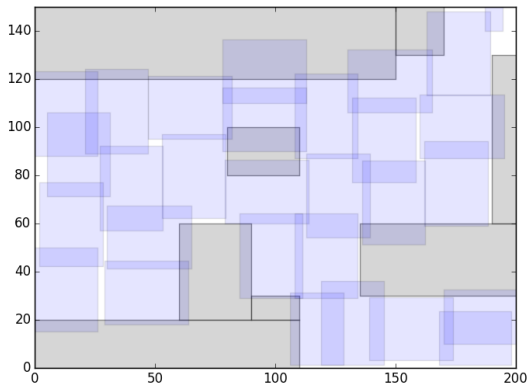
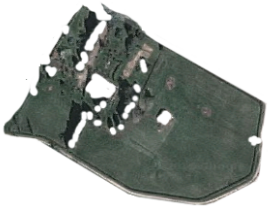
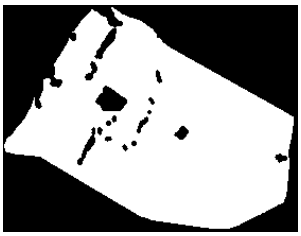


Fig. 8. Cameras positioning with a set of 29 cameras to cover 97.92% of the area.



(a) GPS image of the area to area



(b) Mask of the area to cover

(c) Area cover by the network of cameras

Fig. 9. Coverage area from satelit image with 35 cameras for 86.3% of coverage

REFERENCES

[1] Chin, W. Ntafos, S.: Optimum watchman routes. Inform. Process. Lett. (1988)

[2] Moeini, M., Krller, A., Schmidt, C. Une Nouvelle Approche Pour la Resolution du Probleme de la Galerie dArt.

[3] Erdem, U. M., & Sclaroff, S. (2006). Automated camera layout to satisfy task-specific and floor plan-specific coverage requirements. Computer Vision and Image Understanding, 103(3), 156-169.

[4] Packer, E. (2008). Computing multiple watchman routes. In Experimental Algorithms (pp. 114-128). Springer Berlin Heidelberg.

[5] Zhao, J., Cheung, S. C., & Nguyen, T. (2008). Optimal camera network configurations for visual tagging. Selected Topics in Signal Processing, IEEE Journal of, 2(4), 464-479.

[6] Song, B., Soto, C., Roy-Chowdhury, A. K., & Farrell, J. (2008, September). Decentralized camera network control using game theory. In Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on (pp. 1-8). IEEE.

[7] Liu, L., Zhang, X., & Ma, H. (2010). Optimal node selection for target localization in wireless camera sensor networks. Vehicular Technology, IEEE Transactions on, 59(7), 3562-3576.

[8] Ma, H., Yang, M., Li, D., Hong, Y., & Chen, W. (2012, March). Minimum camera barrier coverage in wireless camera sensor networks. In INFOCOM, 2012 Proceedings IEEE (pp. 217-225). IEEE.

[9] Wang, Q., Wu, J., & Long, C. (2013, October). On-line configuration of large scale surveillance networks using mobile smart camera. In Distributed Smart Cameras (ICDSC), 2013 Seventh International Conference on (pp. 1-6). IEEE. Transactions on, 2004, vol. 52, no 3, p. 771-779.

[10] Zhou, P., & Long, C. (2011, October). Optimal coverage of camera networks using PSO algorithm. In Image and Signal Processing (CISP), 2011 4th International Congress on (Vol. 4, pp. 2084-2088). IEEE.

[11] Reddy, K. K., & Conci, N. (2012, October). Camera positioning for global and local coverage optimization. In Distributed Smart Cameras (ICDSC), 2012 Sixth International Conference on (pp. 1-6). IEEE.

[12] Boeringer, Daniel W. et Werner, Douglas H. Particle swarm optimization versus genetic algorithms for phased array synthesis. Antennas and Propagation, IEEE

[13] [13] Shi, X. H., Liang, Y. C., Lee, H. P., et al. An improved GA and a novel PSO-GA-based hybrid algorithm. Information Processing Letters, 2005, vol. 93, no 5, p. 255-261.

[14] http://www.coppeliarobotics.com/

[15] SRINIVAS, Mandavilli et PATNAIK, Lalit M. Adaptive probabilities of crossover and mutation in genetic algorithms. Systems, Man and Cybernetics, IEEE Transactions on, 1994, vol. 24, no 4, p. 656-667.

[16] MILLER, Brad L. et GOLDBERG, David E. Genetic algorithms, tournament selection, and the effects of noise. Complex Systems, 1995, vol. 9, no 3, p. 193-212.

[17] EBERHART, Russ C., KENNEDY, James, et al. A new optimizer using particle swarm theory. In : Proceedings of the sixth international symposium on micro machine and human science. 1995. p. 39-43.

[18] HOLLAND, John H. Outline for a logical theory of adaptive systems. Journal of the ACM (JACM), 1962, vol. 9, no 3, p. 297-314.

Shape from polarization in the far IR applied to 3D digitization of transparent objects

by A. Zanzouri Kechiche*, R.Rantoson**, O. Aubreton*, F.Meriaudeau***, and C.Stolz*

* LE2I UMR6306, CNRS, Arts et Métiers, Univ. Bourgogne Franche-Comté, 12 Rue de la fonderie, Le Creusot, 71200, France,
** ISIT UMR 6284, CNRS, Faculté de médecine, 28 place Henri Dunant 63 001 Clermont-Ferrand France
***Centre for Intelligent Signal and Imaging Research (CISIR) Electrical & Electronic Engineering Department Universiti Teknologi Petronas 32610 Seri Iskandar, Perak, Malaysia
abir.kechiche@u-bourgogne.fr

Abstract

This paper presents a new application of "shape from polarization" method in the far Infrared range with applications for three-dimensional reconstruction of transparent objects. Shape from polarization is a recent application of more general polarization imaging technique having the aim to digitize the shape of the observed object. The principle is to evaluate the normal on each observed point followed by an integration procedure. The technique is well developed in the visible domain, but not in the far infrared domain due to the requirement of telecentric optics. We propose here a complete setup in the 8-13 micrometer spectral band with an appropriate source and a reconstruction method including the pinhole camera model in order to use standard optics of the camera. We present three-dimensional digitization of transparent objects.

Keywords: *Polarization, thermal imaging, transparent objects, long wave infrared and three-dimensional digitization*

1. State of the art

Recovering three-dimensional surface shape of objects, classified as shape form polarization technique, is one of a classic and important research areas of computer vision. This approach is based on a special property of the light reflected from the object: its polarization state. This is an unconventional approach because it does not rely on the parameters traditionally operated machine vision such as the intensity and wavelength. A standard camera model limited by brightness and hue is not enough to measure the polarization parameters. However, when dealing with the inspection of transparent or highly reflective surfaces like mirror, standard solutions are not available yet. Referring to these non "Lambertian" surfaces, Ihrke et al.[1] published an exhaustive survey which was recently completed by Meriaudeau et al.[2]for transparent objects.

Among the recent works, approaches such as "scanning from heating"[3], where the three-dimensional data is reconstructed from a heat pattern visualized with a calibrated IR camera or "shape from induced fluorescence" [4] where the three-dimensional reconstruction takes place thanks to a generated visible pattern induced by florescence onto the object surface, appear to be very promising for three-dimensional inspection with potential adaptation for transparent as well as specular objects.

In this paper, we used the shape from polarization method in the far infrared band. This spectral band corresponds to a part of the spectrum where transparent materials like glass and some plastics appear opaque. Since the initiator works of Wolff [5] on polarization imaging, recent extensions are made possible by active lightings systems [6],multispectral approach [7] and infrared imaging. In the field of remote sensing Middle IR band is then used for material detection and separation [8]. Another example used a polarimetric camera to record the stokes parameters and the degree of linear polarization of far infrared radiation emitted by human faces to get three-dimensional facial image [gurton]. The last example used the near IR band [9] in order to build a depth map of the outdoor environment. This brings new interest in polarization imaging techniques.

2. Proposed Method

The acquisition and reconstruction procedure is as follows (Figure1): acquisition of an image sequence by rotating a linear polarizer, evaluation of the Stokes parameters, estimation of the normal and of the 3D surface.

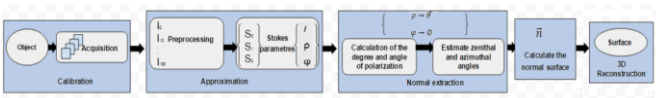


Figure 1: Schematic summary of the proposed reconstruction process

Based on the fact that after reflection, a non-polarized light wave becomes partially linearly polarized, the aim of the shape from polarization is to measure the normal at each

observed point and then to obtain the whole surface by integration of the normals field as in shape shading techniques[10].

A partially polarized wave may be defined by three-parameters which are: the light intensity I, the degree of polarization ρ and the angle of polarization φ. Each of these parameters can also be defined with the stokes parameters S_0, S_1 and S_2 [11].Since no ellipticity is measured, a partial Stokes sensor with a rotative polarizer (figure 1) or equivalent with electro-optics components is sufficient. The intensity of the wave impinging on the camera is given by Eq.1.

$$\begin{cases} I(\alpha) = \frac{1}{2}(s_0 + s_1 \cos 2\alpha + s_2 \sin 2\alpha) \\ I(\alpha) = \frac{I}{2}(1 + \rho \cos(2\alpha - 2\phi)) \end{cases} \quad (1)$$

ρ and φ are respectively linked to the zenithal angle θ and to the azimuthal angle ϕ through the Snell-Descartes relation. These measurements enable to infer the normal at each point as recalled by Eq.2.

$$\vec{n} = \begin{Bmatrix} -\frac{\partial f(x,y)}{\partial x} \\ -\frac{\partial f(x,y)}{\partial y} \\ 1 \end{Bmatrix} = \begin{cases} p = \tan\theta \cos\phi \\ q = \tan\theta \sin\phi \\ 1 \end{cases} \quad (2)$$

3. Experimental Set-up

The common procedure for evaluating the polarization parameters begins by evaluating the three stokes parameters. This requires a sequence of images, by rotating the polarizer with a constant step of 10 degrees, and perform least square fitting and finally evaluate the degree and angle of polarization. However, as pointed out earlier by Miyazaki[12] and Morel[6], evaluation of $\vec{n}$ by these measures are not straightforward since both of them provides two candidates. That's why the use of an active light source was introduced by Morel[6] who calculates an image mask Iquad in order to property solve the ambiguity for the azimuthal angle in the case of metallic surfaces. Concerning the zenithal angle, Miyazaki[13] introduced a multispectral technique considering visible and IR measurement. This needs two different setups, one in the visible and one in the IR, implying either registration or a cumbersome optical system. Also one of the main constraints of standard shape from polarization setups is the use of orthographic projection that is physically realized by telecentric optics. To overcome these two problems, we propose an active shape from polarization experiment in the far Infrared band.

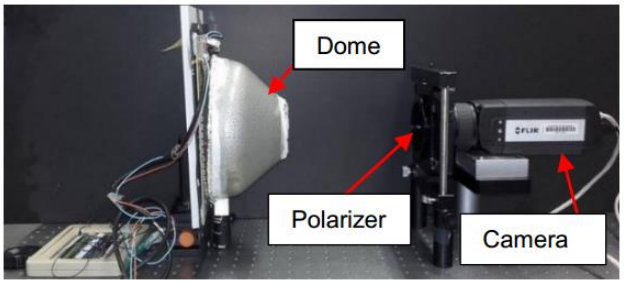


Figure 2: Experimental Set-up

This includes (Figure 2) a thermal camera (Flir A645) of [8 µm -13 µm] range, of 640x480 resolution, a non telecentric 24.5mm focal lens, a manual rotating (ZnSe) polarizer of [1 µm-15 µm] range with an orientation angle α. Also a specific dome (generating the quadrant active IR lighting) is composed of two pieces: a metallic cover which top is holed to place the camera, and a slab of 56 resistor (Figure 3) (12 Ohms and 0.25W each) supplied by a voltage of 12V for each quadrant illumination. The temperature of each resistor is constant, ~60 degrees Celsius, during the acquisitions, corresponding to a maximum radiation of 8.7 µm according to Wien's law.

Figure 3: Lighting system in the IR made of resistors and which can be turned on by quadrants

Since we use the standard lens of the camera, the projection model is now the perspective model, so we choose to be in conditions as close as possible to be original orthographic projection. The pinhole lens model is recalled by Eq.3 for the transformation between 3D coordinates (u, v, z) and the image coordinates (u, v) where α_u and α_v characterize the focal distance, u_0, v_0 the position of the optic center of the camera and s is the scale factor.

$$s\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} \alpha_u & 0 & u_0 & 0 \\ 0 & \alpha_u & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \quad (3)$$

In this model all the rays are converging to the optical center of the camera. That's why we use a relaxed perspective model where the distance camera-object Z_0 is

sufficiently far from the camera [14]. We use the virtual plane at this distance as the basis of the reconstruction.

The method requires to calibrate the camera in order to estimate the appropriate distance Z_0. We used the well-known Zhang-Bouguet method [15]. This requires the use of a checkerboard in order to acquire a sequence of images at different depth and orientation that permits to extract a set of world points necessary for estimating the intrinsic parameters. In our case, the checkerboard was realized with the etching technology of electronic board, so half of the squares are in cupper and the order in the resin material each of them having a different emissivity in the Infrared (Figure 4a).due to the high sensitivity of the camera, no important heating is needed on the checkerboard as shows the raw thermal image of (figure 4b). Notice that for better accuracy in the future processing, all the images are pre-processed (Figure 4c) to grayscale and to have high contrast.

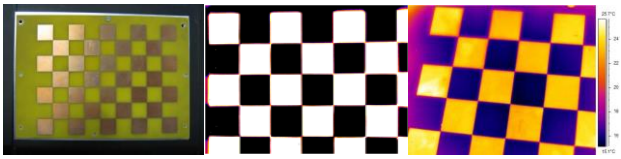


Figure 4: a) View of the checkerboard made of two materials of different emissivity in IR b) Raw thermal image. c) Processed view of the thremal image in graylevels

Once this calibration is done, we can determine the appropriate scale factor s by using two reference points: one on the object and one on the support of the object. Z_0 is defined as the mean distance between the reference points multiplied by the scale factor s. In our experiments Z_0 was estimated to 300 mm.

The other consequence of the perspective model will be the non-regularity of the (x,y) coordinates grid compared to the former orthographic model. That's why we adapted the integration method by using the successive Over Relaxation approach [16].

4. Results

To sum up, the acquisition and reconstruction procedure is as follows: acquisition of the image sequence by rotating the polarizer, acquisition of the mask image [6] by sequentially powering the resistors and evaluation of the Stokes parameters with verification of their accuracy in order to remove noisy data points. The criterion evaluates the physical admissibility of the parameters, $S_0 \geq \sqrt{S_1^2 + S_2^2}$, completed by the comparison between the measured intensity I by the sensor and the approximated intensity I: we consider the relative error between I and I, the sign of the product of their respective derivative and the relative error of these derivatives [14]. This admissibility criterion is visualized as binary image where black pixels satisfy the criterion. If more than 80% of pixels are satisfactory (Figure 5b), we calculate the polarization parameters (Figure 6a and Figure 6b), the zenithal and azimuthal angles after the normals field (Figure 7) and so the 3D shape.

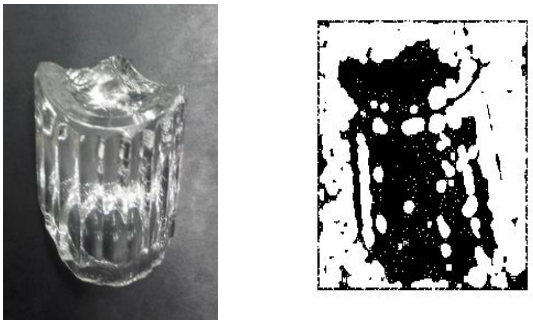


Figure 5: a) Glass object b) Quality map for the evaluated Stokes parameters (black pixels are good)

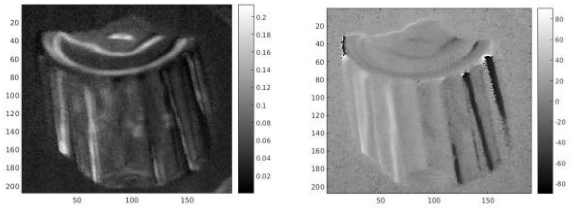


Figure 6: a) Degree of polarization, b) Angle of polarization

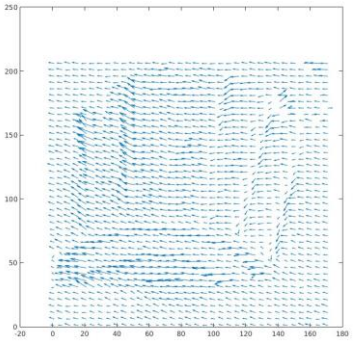


Figure 7: The normals field

Furthermore, the dispersion and absorption of the far IR light is due to impurities in the glass matrix. This implies a complex value of the refraction index that shifts the value of the Brewster angle to high values, enabling to infer the correct value of the zenithal angle θ from the degree of polarization for surface with angle ranging from 0 to maximum values close to 80°. However, even for the Brewster angle, the degree of polarization is less than 1.

This can be compensated by introducing a pseudo-index, where n is the real index and k the coefficient of extinction of the material being studied. This value has to be estimated empirically for each studied sample. This is done by choosing an index so that the estimated value of the Brewster angle matches the maximum value of the measured degree of polarization. For example, the curve of Figure .8 corresponds to a pseudo-index estimated to 1.6+3.5i for the object presented in Figure 5.a. The value of 0.27 at the Brewster angle is closed to the measured degree of polarization of 0.24 (Figure 6.a). It has the advantage to take care of the noise and finely tune the final shape.

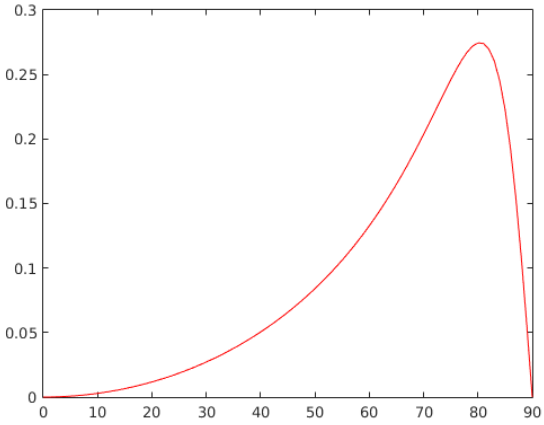


Figure 8: Evolution of the degree of polarization versus the zenithal angle with evaluated index of $1.6+3.5i$ (reel and theory graph).

Our experiments were first conducted with the glass object of Figure 5.a that is reconstructed by following the previous steps. As we see in Figure 9.a and Figure 9.b the reconstructed shape is in close agreements with the observed object (Figure5.a).

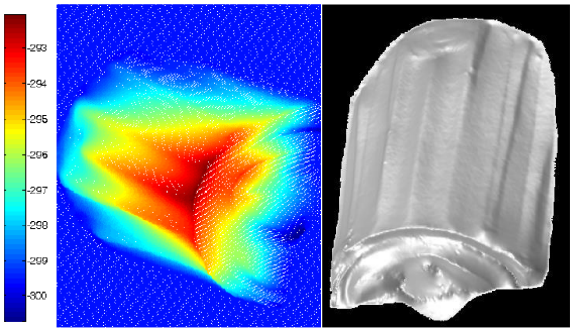


Figure 9: 3D reconstruction of the tested glass object (a) matlab and (b) rapidform result

5. Conclusion

In conclusion, we presented an extension of the shape from polarization within the long Infrared range. The active lighting setup as well as the use of long IR range enables us to remove the ambiguities on both the zenithal angle and azimuthal angle. Also the extension of the reconstruction model by using the perspective lens model permits us to avoid the use of costly telecentric lens. Promising preliminary results were obtained and further experiments are still being carried out to further validate the accuracy of the system. A second ring of resistors can then be added to the IR source in order to apply a multispectral approach [17].

References

[1] I. Ihrke, K. N. Kutulakos, H. P. A. Lensch, M. Magnor, and W. Heidrich, "Transparent and Specular Object Reconstruction", Computer Graphics Forum 29, 2400-2426 (2010).

[2] F. Meriaudeau, R. Rantoson, D. Fofi, and C. Stolz, "Review and comparison of Non Conventional Imaging Systems for 3D Digitization of transparent objects", Journal of Electronic Imaging 21, 021105 (2012).

[3] F. Meriaudeau, L. A. Sanchez-Secades, G. Eren, A. Erçil, F. Truchetet, O. Aubreton, and D. Fofi, "3D Scanning of Non-Opaque Objects by means of Infrared Imaging", IEEE Transaction on Instrumentation and Measurement 59 2898-2906 (2010).

[4] R. Rantoson, C. Stolz, D. Fofi, and F. Meriaudeau, "Optimization of transparent objects digitization from visible fluorescence ultraviolet induced", Opt. Eng. 51, 033601-033601 to 033601-033610 (2012).

[5] L. B. Wolff, "Polarization vision: a new sensory approach to image understanding", Image and Vision computing 15, 81-93 (1997).

[6] O. Morel, C. Stolz, F. Meriaudeau, and P. Gorria, "Active Lighting Applied to 3D Reconstruction of Specular Metallic Surfaces by Polarization Imaging", Appl. Opt. 45, 4062-4068 (2006).

[7] M. Ferraton, C. Stolz, and F. Mériaudeau, "Optimization of a polarization imaging system for 3D measurements of transparent objects", Opt. Express 17, 21077-21082 (2009).

[8] F. Cremer, W. d. Jong, K. Schutte, W.-J. Liao, and B. A. Baertlein, "Detectability of surfacelaid landmines with a polarimetric IR sensor," in Detection and Remediation Technologies for Mines and Minelike Targets VIII, (SPIE, 2003).

[9] F. A. Sadjadi, "Passive three-dimensional imaging using polarimetric diversity", Opt. Lett. 32, 229-231 (2007).

[10] R. Zhang, P. S. Tsai, J. E. Cryer, and M. Shah, "Shape from shading, A survey", IEEE Trans. Pattern Anal. Mach. Intell. 21, 690-796 (1999).

[11] D. L. Goldstein, Polarized light second Edition (Optical Engineering) (CRC, 2003), p. 680.

[12] D. Miyazaki, M. Kagesawa, and K. Ikeuchi, "Determining Shapes of Transparent Objects from Two Polarization Images," in IAPR, (IEEE, 2002), 26-31.

[13] D. Miyazaki, M. Saito, Y. Sato, and K. Ikecuhi, "Determining Surface Orientations of transparent objects on polarization degrees in visible and infrared wavelength", J. Opt. Soc. Am. A 19, 687-694 (2002).
[14] R. Rantoson, C. Stolz, D. Fofi, and F. Meriaudeau, "3D Reconstruction by polarimetric imaging method based on perspective model," in Europe Optical Metrology, (SPIE, 2009).
[15] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in International Conference on Computer Vision, (IEEE, 1999), 666.
[16] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, Numerical Recipes in C++, 3rd edition ed. (Cambridge University Press, 2007).
[17] C. Stolz, M. Ferraton, and F. Meriaudeau, "Shape from polarization: a method for solving zenithal angle ambiguity", Opt. Lett. 37(2012).