Université Bourgogne Franche-Comté - Doctoral Day 2019

Collaborative Writing and Publishing by ImViA Students

Presented by ERL VIBOT students team

Abir Zanzouri, Ahmad Zawawi Jamaluddin, David Strubel, Daniel Braun, Marc Blanchon, Thibault Clamens, Thomas Herrmann and Yifei Zhang

Contents

Doctoral day SPIM Vol: 4 (2019)

A deep learning approach for the segmentation of myocardial diseases

Khawla Brahim[1,2,3], Arnaud Boucher[2], Anis Sakly[3], and Fabrice Meriaudeau[2]

[1]National Engineering School of Sousse, University of Sousse, Sousse, Tunisia
[2]ImViA EA 7535 laboratory, University of Burgundy, Dijon, France
[3]LAESE laboratory, National Engineering School of Monastir, University of Monastir, Monastir, Tunisia

***Abstract*—We present a new multifield expert annotated dataset that can be used to train existing deep learning architecture for automatic segmentation of myocardial disease (infarct core and no reflow region) in LGE-MRI (Late gadolinium enhanced magnetic resonance imaging) with highly robust and reproducible results. Two state-of-the-art medical deep network are compared using our dataset. Given that the scar tissue represents a small part of the whole MRI slices, only myocardium area was considered in the input patches. We show that this preprocessing step facilitate the learning procedure. The final network segmentation performances will be useful for futur comparison of new method to the current related work for this task. Experiments over our proposed dataset, using several evaluation metrics such as mean absolute error, Fscore and area under the ROC curve measures, show the efficiency of these two recent state-of-art methods in quantifying different zones of myocardium infarction. More interestingly, on training infarcted tissue, Unet and Lungnet methods yielded an overall mean absolute error, respectively, of 0.0079372 and 0.067627.**

***Index Terms*—deep learning architecture, segmentation, myocardial disease**

I. Introduction

According to the World Heart Organization (WHO) [1], myocardial Infarction (MI) is a severe silent killer in the world. Currently, there is a great demand for automatic quantifying diseased myocardial slices. The use of magnetic resonance contrast agents based on gadolinium-chelates for visualizing the scarred myocardium is considered as the most clinical relevent references for better myocardial infarction (MI) diagnosis. Biological studies demonstrated that enhancement on late gadolinium-enhanced (LGE) MRI, as a consequence of the accumulation of the agent in the damaged tissue, shows infarcted myocardium, whereas non enhancement indicates the presence of viable myocardium. Infarcted tissue may also present hypointense regions due to the permanent microvascular obstruction (MVO, also call no reflow) phenomenon. MVO indicates the lack of reperfusion of some myocardial region even after the ending of the ischemic event [2].

In this work, Unet [3] and Lungnet [4] methods were tested for detection and quantification of myocardial infarction from short-axis cardiac LGE-MRI. The algorithm provides the left-ventricle myocardium segmentation, per-slice detection of diseased heart and quantification of damaged myocardial areas.

This paper is organized as follows. Related work are briefly reviewed in Sect. II. Materials and Methods are explained in Sect. III. A comparative experimental study is described in Sect. IV and the conclusion is given hereafter.

II. Related work

A. Reference Datasets on Cardiac LGE-MRI

Techniques reproducibility and comparison against other methods are mainly constrained to the private scans used in the studies. Under the STACOM 2012 challenge a first attempt of a reference cardiac LGE-MRI dataset was presented accounting with pathological human (15 cases) and animal (15 cases) data [5], [6]. Experts annotations were centered in hyper-enhanced tissue regions delineation but without providing no reflow phenomena ground truth.

B. Myocardial Infarction Segmentation

Myocardial Infarction Segmentation is interested in effectively Quantifying myocardial scar. In this context, the most frequently used works are the threshold based ones, such as the full-width at half-maximum (FWHM) [7] and the n-standard deviations (from now n-SD) [8]. Nevertheless, these techniques provide significant differences with expert delineations and high result variability [9], [10]. Common several studies recombined the thresholding methods [11]–[13] or used intensity features with connected component analysis [14]–[16] or SVM [17], [18]. Graph-cuts [19], [20] and watershed algorithms [21], [22] have also acheived researchers interest. The 2D U-Net convolutional-neural-network (CNN) winner of the ISBI 2012 challenge [3] showed outstanding performances for segmenting the myocardium on kinetic MRI under the ACDC challenge. [23].

III. Materials and Methods

A. Purpose

In this work, we sought to detect changes of enhanced and non enhanced tissue on late gadolinium-enhanced magnetic resonance imaging in patients with a myocardial infarction (MI). Infarcted myocardial tissue may present hypointense areas as a result of the permanent microvascular obstruction (MVO) phenomenon Fig. 1. We exploit recent state-of-the-art deep learning architecture [3], [4] for automatic detection of viable myocardiac segments.

The architecture of Unet [3] and Lungnet [4] methods are respectively shown in Fig. 2 and Fig. 4.

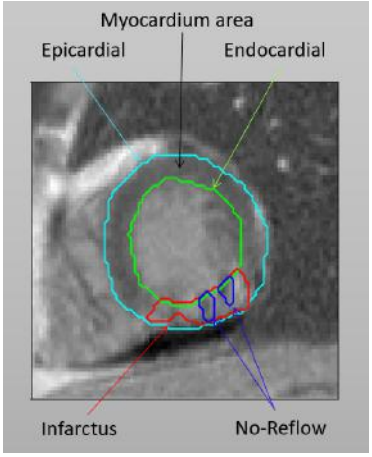


Fig. 1. short-axis view of LGE-MRI

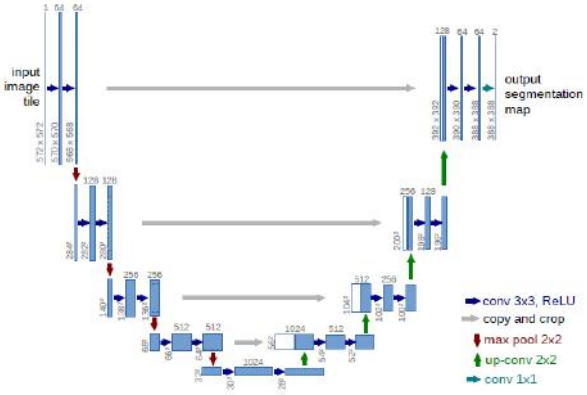


Fig. 2. U-net architecture (example for 32x32 pixels in the lowest resolution) [3]

B. Data annotation

A cohort of healthy and myocardial infarcted patients which attended the imaging center of the University Hospital of Dijon (Dijon, France) between 2015 and 2017 were included in the work. We collected a large expert annotated LGE-MRI Dataset that will be opened in a public repository. The full database accounts with 751 images from which 474 presents diseased myocardium (111 including microvascular obstructions) and 277 healthy. The dataset ground truths were drown in each slice by an expert of the institution (AL) with more than 10 years of expertise in the field. Manually drawn delineations are given for endocardial, epicardial, infarcted myocardium and microvascular obstruction regions. All datasets were stored using the digital imaging and communications in medicine (DICOM) format and anonymized for research purposes.

IV. Experimental Results

A. Performance evaluation criteria

In order to improve the results in this method, we only consider the myocardium area in the input patches. Representative images shown in Fig. 5.

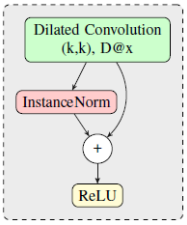


Fig. 3. The block function of the proposed architecture. (k, k) is the size of the convolution kernel and x is the dilation rate [4]

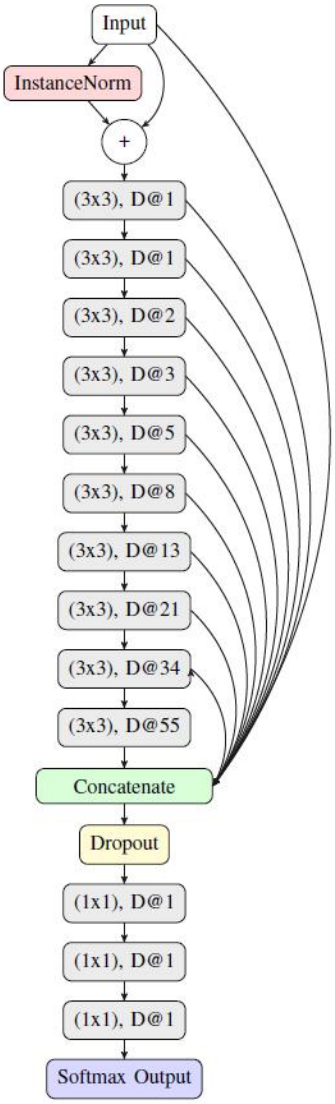


Fig. 4. The architecture of Lungnet [4]. Each gray box corresponds to a block like the one presented in Fig. 3

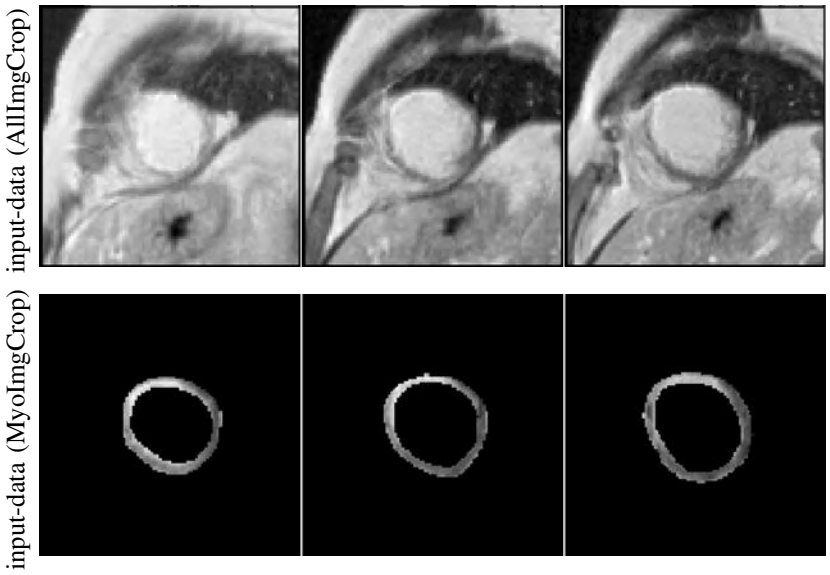


Fig. 5. Three slices from three different input patients. From top to bottom : input-data (AllImgCrop : the original image cropped around the area of interest), input-data (MyoImgCrop : the myocardium zone cropped around the area of interest)

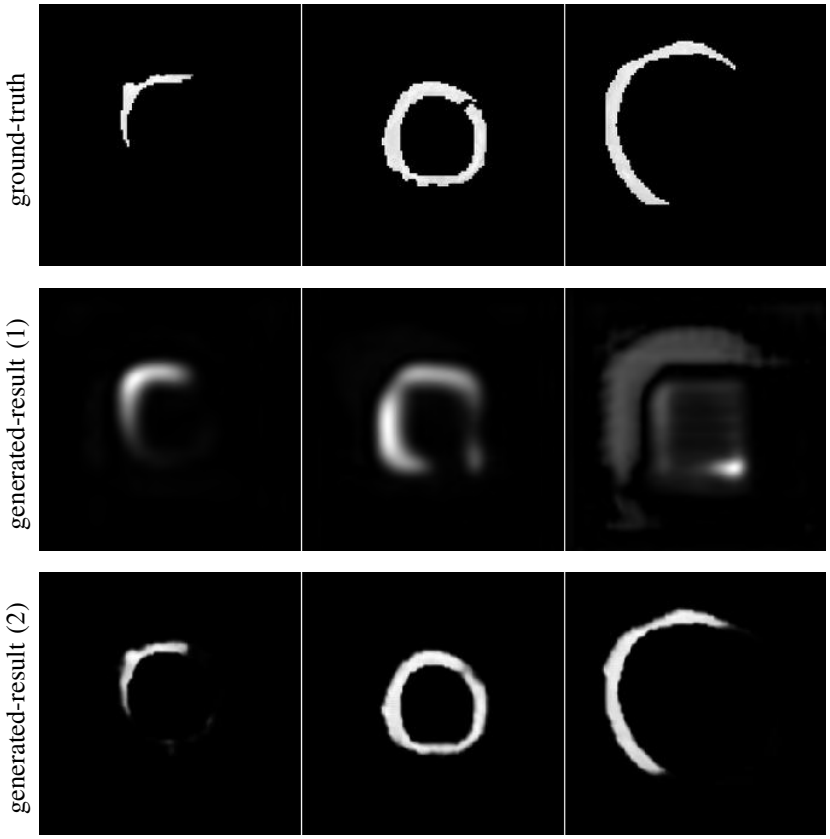


Fig. 6. Elucidation of unet method steps. From top to bottom : ground truth, generated-result (1) (input-data (AllImgCrop : the original image cropped around the area of interest)), generated-result (2) (input-data (MyoImgCrop : the myocardium zone cropped around the area of interest))

We used five types of metrics to measure the performance of the deep learning model: (1) the Mean Absolute Error (MAE); (2) the F-score; (3) the Precision-recall (PR); (4) the Receiver operating characteristic (ROC) and computing its (5) Area Under Curve.

The Precision (P) gives the percentage of correct myocardium pixel-based detected. It is defined as the ratio of the generated output correctly detected to ground truth, so that an ideal result corresponds to the highest accuracy (1)

$$P = \frac{\sum_{x,y} Res(x,y).GT(x,y)}{\sum_{x,y} Res(x,y)} \quad (1)$$

The Recall (R) measures the quantity of ground truth pixels detected (GT) in the generated output (2).

$$R = \frac{\sum_{x,y} Res(x,y).GT(x,y)}{\sum_{x,y} GT(x,y)} \quad (2)$$

The F-Score measure [24] is adopted in order to evaluate the performance of the generated output from the notions defined previously and it is given by the (3). Therefore, an the perfect segmentation approach corresponds to a F-Score measure of 1.0.

$$F_\beta = \frac{(1+\beta^2).P.R}{\beta^2.P+R} \quad (3)$$

We set β^2 at 0.3 to increase the precision influence.

The Mean Absolute Error (MAE) provides a new means of evaluation, which directly quantifies how identical a generated output is to the ground truth (GT). Then MAE is defined as the eq. (4).

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{H} \sum_{y=1}^{W} |Res(x,y) - GT(x,y)| \quad (4)$$

where H and W are respectively the height and width of the image.

Receiver Operating Characteristics Curve [25] plots the false positive rate (FP) representing the pixels scar but detected healthy by the deep learning algorithm while varying a consistent threshold from 0 to 255 against the truth positive rate (TP) which are the healthy pixels detected correctly. An ideal method has a 0% value of (FP) and 100% value of (TP).

Finally, we computed the Area Under the ROC Curve (AUC) which shows how well the proposed method portends against the ground truth (GT), $(0 < AUC < 1)$. A perfect prediction leads to a higher AUC area.

B. Results

The experimental result shows that Unet [3] and Lungnet [4] models can obtain good localization of the myocardial area. Our results were compared to the ground truth marked by the expert cardiologist. Fig. 6 illustrates that Unet [3] method can also accurately detect the MI area. Table I summarizes the diverse values of F-score over two state-of-the-art deep learning models on CHU-Dijon dataset. AUC area presented in Table II gives a more promising result. The area under

TABLE I
RESULTS. F-SCORE VALUES

Methods	Myocardial Area	Myocardial Infarction
UnetTrain-AllImgCrop	0.8191	0.2658
UnetTest-AllImgCrop	0.7296	0.1616
LungnetTrain-AllImgCrop	0.5643	0.1823
LungnetTest-AllImgCrop	0.6517	0.1711
UnetTrain-MyoImgCrop	0.9624	0.5388
UnetTest-MyoImgCrop	0.9630	0.4752
LungnetTrain-MyoImgCrop	0.9162	0.4864
LungnetTest-MyoImgCrop	0.9432	0.4604

TABLE II
RESULTS. AUC VALUES

Methods	Myocardial Area	Myocardial Infarction
UnetTrain-AllImgCrop	0.9843	0.5140
UnetTest-AllImgCrop	0.9534	0.4203
LungnetTrain-AllImgCrop	0.9347	0.4935
LungnetTest-AllImgCrop	0.9609	0.4588
UnetTrain-MyoImgCrop	0.9995	0.5690
UnetTest-MyoImgCrop	0.9993	0.5537
LungnetTrain-MyoImgCrop	0.9982	0.6512
LungnetTest-MyoImgCrop	0.9987	0.6284

the ROC curve and F-score values have being significantly improved by cropping input image to be centered to the area of interest (Myocardium zone). As shown in Fig. 7 and Fig. 8 the MAE values for predicting infarct and myocardial areas segmentation were very small on both current deep learning models. The ROCs and PRs curves are shown in Fig. 9, Fig. 10, Fig. 11 and Fig. 12. These results indicate the potential of Unet [3] and Lungnet [4] frameworks in detecting damaged myocardial areas. The extensive validation of both existing methods on our new benchmark turns this proposal into a robust tool with clinical transfer potential.

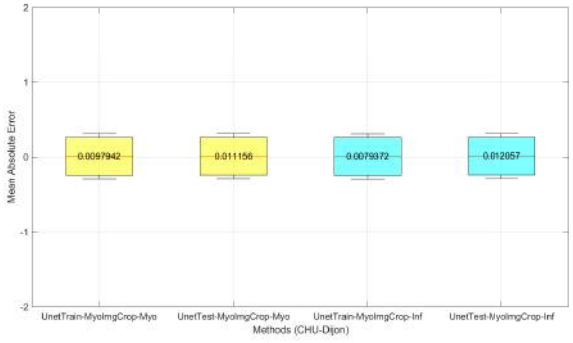


Fig. 7. MAE values (Unet)

In this paper, we present an end-to-end algorithm for myocardial infarction segmentation and quantification in LGE-MRI. It is interesting to measure myocardial infarct size in gadolinium-enhanced magnetic resonance imaging and to validate these methods in detecting the accurate location of a

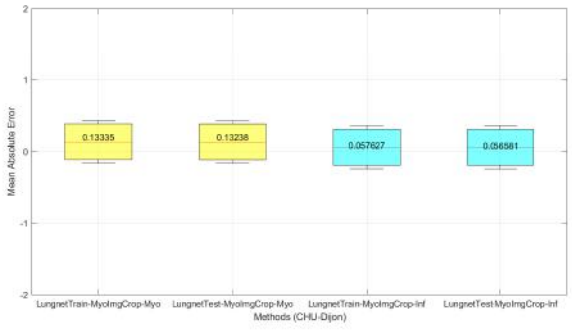


Fig. 8. MAE values (Lungnet)

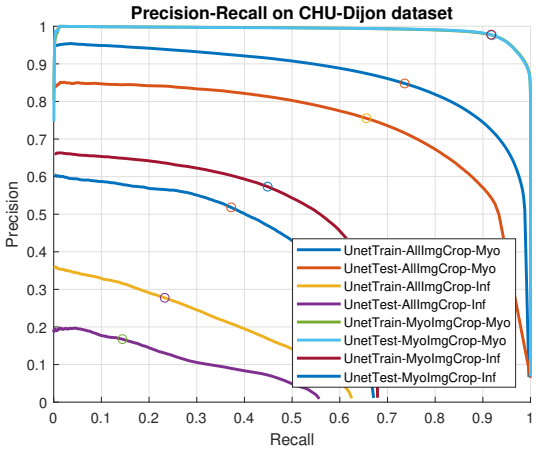


Fig. 9. Precision-Recall curves on CHU-Dijon dataset (Unet)

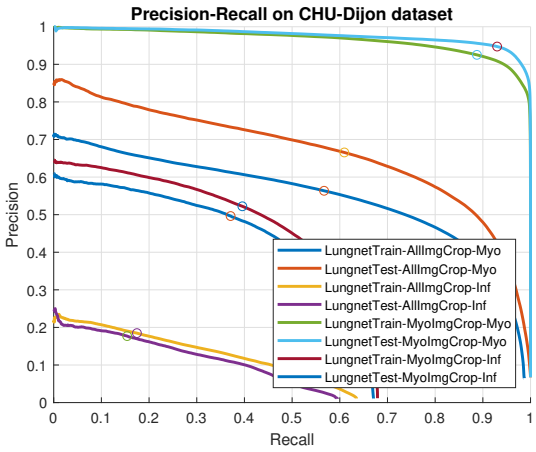


Fig. 10. Precision-Recall curves on CHU-Dijon dataset (Lungnet)

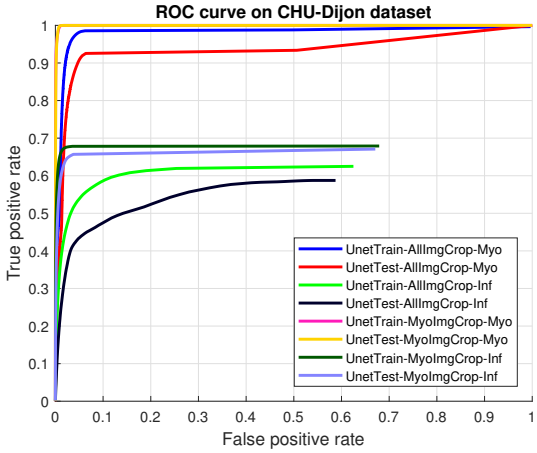


Fig. 11. Precision-Recall curves on CHU-Dijon dataset (Unet)

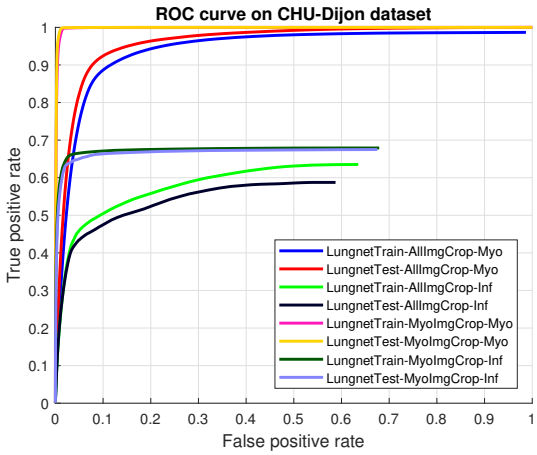


Fig. 12. ROC curves on CHU-Dijon dataset (Lungnet)

myocardial areas. A future investigation is intended to exploit the strength of each medical deep learning segmentation model.

REFERENCES

[1] Organization, W.H., Organization, W.H., et al.: The atlas of heart disease and stroke. In World Health Organization: Geneva. (2004).

[2] Rajiah, P., Desai, M. Y., Kwon, D., Flamm, S. D. : MR imaging of myocardial infarction. In Radiographics, 33(5), pp. 383–1412. (2013).

[3] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241. (2015).

[4] Anthimopoulos, M., Christodoulidis, S., Ebner, L., Geiser, T., Christe, A., Mougiakakou, S.: Semantic Segmentation of Pathological Lung Tissue with Dilated Fully Convolutional Networks In arXiv preprint arXiv:1803.06167 (2018).

[5] Karim, R., Bhagirath, P., Claus, P., Housden, R. J., Chen, Z., Karimaghaloo, Z., Sohn, H.M., Rodríguez, L.L., Vera, S., Albà, X., et al., A.: Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late Gadolinium enhancement MR images. In Medical image analysis, 30, pp. 95–107. (2016).

[6] Karim, R., Claus, P., Chen, Z., Housden, R.J., Obom, S., Gill, H., Ma, Y., Acheampong, P., ONeill, M., Razavi, R., et al.: Infarct segmentation challenge on delayed enhancement mri of the left ventricle. In Interna-

tional Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 97–104. (2012).

[7] Amado, L.C., Gerber, B.L., Gupta, S.N., Rettmann, D.W., Szarf, G.,Schock, R., Nasir, K., Kraitchman, D.L., Lima, J.A.: Accurate and objective infarct sizing by contrast-enhanced magnetic resonance imaging in a canine myocardial infarction model. In Journal of the American College of Cardiology 44, pp. 2383–2389. (2004).

[8] Kim, R.J., Fieno, D.S., Parrish, T.B., Harris, K., Chen, E.L., Simonetti, O., Bundy, J., Finn, J.P., Klocke, F.J., Judd, R.M.: Relationship of mri delayed contrast enhancement to irreversible injury, infarct age, and contractile function. In Circulation 100, pp. 1992–2002. (1999).

[9] Spiewak, M., Malek, L.A., Misko, J., Chojnowska, L., Milosz, B., Klopotowski, M., Petryka, J., Dabrowski, M., Kepka, C., Ruzyllo, W.: Comparison of different quantification methods of late gadolinium enhancement in patients with hypertrophic cardiomyopathy. In European journal of radiology 74, e149–e153. (2010).

[10] Zhang, L., Huttin, O., Marie, P.y., Felblinger, J., Beaumont, M., Chillou, C.D., Girerd, N., Mandry, D.: Myocardial infarct sizing by late gadolinium-enhanced mri: Comparison of manual, full-width at half-maximum, and n-standard deviation methods. In Journal of Magnetic Resonance Imaging 44, pp. 1206–1217. (2016).

[11] Andreu, D., Berruezo, A., Ortiz-Pérez, J.T., Silva, E., Mont, L., Borràs,, R., de Caralt, T.M., Perea, R.J., Fernández-Armenta, J., Zeljko, H., et al.: Integration of 3d electroanatomic maps and magnetic resonance scar characterization into the navigation system to guide ventricular tachycardia ablationclinical perspective. In Circulation: Arrhythmia and Electrophysiology 4, pp. 674–683. (2011).

[12] Flett, A.S., Hasleton, J., Cook, C., Hausenloy, D., Quarta, G., Ariti, C., Muthurangu, V., Moon, J.C.: Evaluation of techniques for the quantification of myocardial scar of differing etiology using cardiac magnetic resonance. In JACC: cardiovascular imaging 4, pp. 150–156. (2011).

[13] Schmidt, A., Azevedo, C.F., Cheng, A., Gupta, S.N., Bluemke, D.A., Foo, T.K., Gerstenblith, G., Weiss, R.G., Marbán, E., Tomaselli, G.F., et al.: Infarct tissue heterogeneity by magnetic resonance imaging identifies enhanced cardiac arrhythmia susceptibility in patients with left ventricular dysfunction. In Circulation 115, pp. 2006–2014. (2007)

[14] Hsu, L.Y., Natanzon, A., Kellman, P., Hirsch, G.A., Aletras, A.H., Arai, A.E.: Quantitative myocardial infarction on delayed enhancement mri. part i: Animal validation of an automated feature analysis and combined thresholding infarct sizing algorithm. In Journal of Magnetic Resonance Imaging 23, pp. 298–308. (2006).

[15] Tao, Q., Milles, J., Zeppenfeld, K., Lamb, H.J., Bax, J.J., Reiber, J.H., van der Geest, R.J.: Automated segmentation of myocardial scar in late enhancement mri using combined intensity and spatial information. In Magnetic Resonance in Medicine 64, pp. 586–594. (2010).

[16] Valindria, V.V., Angue, M., Vignon, N.,Walker, P.M., Cochet, A., Lalande, A.: Automatic quantification of myocardial infarction from delayed enhancement mri. In Signal-Image Technology and Internet-Based Systems (SITIS), Seventh International Conference on, IEEE. pp. 277–283. (2011).

[17] Dikici, E., O'Donnell, T., Setser, R., White, R.D.: Quantification of delayed enhancement mr images. In International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 250–257. (2004).

[18] O'Donnell, T.P., Xu, N., Setser, R.M., White, R.D.: Semiautomatic segmentation of nonviable cardiac tissue using cine and delayed enhancement magnetic resonance images. In Medical Imaging : Physiology and Function: Methods, Systems, and Applications, International Society for Optics and Photonics. pp. 242–252. (2003).

[19] Lu, Y., Yang, Y., Connelly, K.A., Wright, G.A., Radau, P.E.: Automated quantification of myocardial infarction using graph cuts on contrast delayed enhanced magnetic resonance images. In Quantitative imaging in medicine and surgery 2(2), pp. 81. (2012).

[20] Wei, D., Sun, Y., Ong, S.H., Chai, P., Teo, L.L., Low, A.F.: A comprehensive 3-d framework for automatic quantification of late gadolinium enhanced cardiac magnetic resonance images. In IEEE Transactions on Biomedical Engineering 60, pp. 1499–1508. (2013a).

[21] Hennemuth, A., Seeger, A., Friman, O., Miller, S., Klumpp, B., Oeltze, S., Peitgen, H.O.: A comprehensive approach to the analysis of contrast enhanced cardiac mr images. In IEEE Transactions on Medical Imaging 27, pp. 1592–1610. (2008).

[22] Kruk, D., Boucher, A., Lalande, A., Cochet, A., Sliwa, T.: Segmentation integrating watershed and shape priors applied to cardiac delayed enhancement mr images. In IRBM 38, pp. 224–227. (2017).

[23] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multistructures segmentation and diagnosis: Is the problem solved?. In IEEE Transactions on Medical Imaging. (2018).

[24] Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. In IEEE Transactions on Pattern Analysis and Machine Intelligence 37(3), pp. 569–582 (2015).

[25] Fawcett, T.: An introduction to roc analysis. Pattern recognition letters 27(8), pp. 861–874 (2006).

A Starting Point for Efficient Human Action Detection

Yu Liu
Laboratoire ImViA
Universite de Bourgogne, Dijon, France
alphadadajuju@gmail.com

Fan Yang
Laboratoire ImViA
Universite de Bourgogne, Dijon, France
fanyang@u-bourgogne.fr

Dominique Ginhac
Laboratoire ImViA
Universite de Bourgogne, Dijon, France
dominique.ginhac@ubfc.fr

Abstract

Analyzing videos of human actions involves understanding the spatial and temporal context of the scenes. Existing state-of-the-art approaches have demonstrated impressive results using Convolution Neural Networks (CNNs). However, most of them operate in a non-realtime, offline fashion and are not well-equipped for the emerging real-world scenarios, such as autonomous driving and public surveillance. In addition, they remain computationally demanding as simpler models tend not to suffice for such complicated tasks. This paper reviews state-of-the-art methods based on CNN for human action detection. We also propose an initial pipeline to address both aspects of action detection accuracy and efficiency by exploiting multiple cues and video data continuity. We partially validate its feasibility on the UCF-101-24 dataset, and lay out rigorous future plans to execute the proposed ideas.

1. Introduction

Human action detection is a key element to video understanding. It has been an active research topic driven by many applications, such as assisted or autonomous driving, unmanned surveillance, and robot vision. These real-world scenarios often not only mandate on-site and real-time automatic interpretation of the scene, but also require robust recognition of events under a restricted power budget.

Moreover, applications such as surveillance in large environments and abnormal behavior detection in public, further demand extensive camera networks and exchange of information among local/central devices. The need to transmit and store redundant video streams imposes bottlenecks for effective analytic activities. To manage this enormous data flow from multiple cameras without network overload, efficient processing and extraction of relevant metadata at the local device become a fundamental system requirement. Instead of raw video streams, transmitting metadata between system components not only can minimize the content to be streamed, but also creates a smarter system within a cooperative framework.

With the recently rising Convolution Neural Network (CNN), single-image object detection has progressed significantly with remarkable results [1][2][3][4]. This motivates researchers to adopt CNN object detectors on action detection. To achieve spatio-temporal detection for action instances, they often link the frame-level detection over time to create spatio-temporal tubes [5][6][7][8]. Handling video frames independently is however sub-optimal as the temporal continuity of videos is not fully exploited. Distinguishing actions from a single frame can be ambiguous. In addition, neglecting the content similarity between successive frames imposes high processing cost and redundancy.

Our work is supported by the H2020 Innovative Training Network (ITN) project ACHIEVE. It aims to converge advanced image sensing technologies, pattern recognition and machine learning for distributed vision applications. In particular, our work focuses on addressing computationally inexpensive human action detection, potentially for embedded video scene understanding. In this paper, we review state-of-the arts on action detection and related topics, describe our proposed pipeline and demonstrate its validity in preliminary experiments.

2. Related work

Thanks to their remarkable results on object detection in images, CNN object detectors have been increasingly adopted for video action detection. Here, we review modern single-image object detectors, followed by their extension in video object detection and finally video action detection.

2.1. Object detection in images

Modern CNN object detectors can be grouped into two families. The first one consists of a two-stage approach popularized by R-CNN [9] and its descendants [1][2][10][11]. These methods first propose object regions from images, and then perform classification and bounding box regression for each region. Following R-CNN, SPP-Net [10] and Fast R-CNN [11] accelerate detection speed by handling the proposed regions on a shared feature map to avoid feeding them to CNN multiple times. Faster R-CNN [1] later introduces a simple, convolutional region proposal network and pre-defined candidate boxes ("anchors"), achieving higher inference speed. R-FCN [2], which is fully convolutional, further pushes feature sharing up to the region classification and regression layers to increase detection speed.

Alternative to the two-stage method, YOLO [3] and SSD [4] also employ the concept of anchors with a single-shot approach. The anchor boxes are directly classified and regressed in a single pass without the intermediate region proposal step. In exchange for minor drops in accuracy, these methods can achieve real-time detections.

2.2. Video object detection

Video object detection has been extensively explored since ImageNet introduced the VID challenge [12]. Many existing methods rely on single-image object detection in the post-processing step. Han et al. [13] propose Seq-NMS which links high-confidence bounding boxes from consecutive frames into tubelets. Kang et al. [14] take into account temporal information by applying a tracker and precomputed optical flows to map detection to nearby frames. Instead of post-processing single-image detection, Feichtenhofer et al. [15] set up a CNN architecture with a multi-task objective, simultaneously handling frame-based detection and across-frame track regression in the network.

Neither of the above methods focus on efficiency. On the other hand, Zhu et al. [16] incorporate R-FCN and FlowNet [17] to propagate deep feature maps from keyframes to nearby non-keyframes via flow fields. This accelerates video object detection as only sparse keyframes need to go through the time-consuming deep feature extractor. In a similar spirit, Liu et al. [18] propagate frame-level information across frames using their proposed recurrent-convolutional architecture. Combined with the fast SSD object detector, their approach demonstrates real-time video object detection on low-powered mobile devices.

2.3. Action detection

Many recent approaches for action detection rely on object detectors trained on action classes. Gkioxari et al. [19] incorporate R-CNNs in a two-stream framework, performing detection on the appearance and motion data separately. The resulting per-frame detections are then linked across the temporal domain based on action predictions and spatial overlaps between consecutive frames. Others apply the popular Faster R-CNN in a similar two-stream framework. Saha et al. [5] fuse appearance and motion cues outside the network based on overlaps between their respective detection scores and boxes. On the other hand, Peng et al. [7] introduce a fusion layer, merging region proposals from both streams into fixed-length feature vectors for action classification and box regression. Targeting real-world application scenarios, Singh et al. [8] achieve action detection in a real-time, online manner by combining the two-stream approach, efficient SSD detectors, fast optical flow estimator [20] and their proposed online linking algorithm. Instead of making detection at the frame level, Kalogeiton et al. [6] propose the action tubelet detector, which learns to output sequences of action bounding boxes and scores, i.e., action tublelets.

Inspired by the two-stream approach, Zolfaghari et al. [21] introduce a Markov chain model which adds appearance, motion and pose cues successively to refine results in the action recognition task. In a similar spirit, Wang et al. [22] leverage semantic cues (e.g. scene, person, object) in addition to the original two-stream CNN for action understanding in videos.

3. Methodology

In this paper, we consolidate previous works and concepts from literature that are relevant to our problem. Our proposed pipeline is depicted in Fig. 1.

3.1. Multi-stream action detection

As outlined in Fig. 1, our approach will adopt the multi-stream CNN framework to model multiple modalities (starting with two modalities: appearance and motion) separately in action videos. CNN object detectors, each independently handling a type of stream, will predict detection boxes and class-specific scores for different video frames. We will experiment with different fusion strategies to aggregate the appearance and motion cues. As suggested in [5][7], some possible methods include concatenating appearance and motion features, or performing box fusion as a post-network operation.

Existing two-stream approaches often capture motion information in the form of optical flow sequences, precomputed using traditional optical flow estimation methods [23]. This is time consuming and storage demanding. Moreover, such optical flows are computed independent of the specific tasks in hand (e.g., action detection). It is therefore our desire to adopt the framework in [24][16] which integrates learning of flow estimation in the network. In addition to frame-level action detection, we will also explore aggregation of features across consecutive frames to capture more temporal information [25][26].

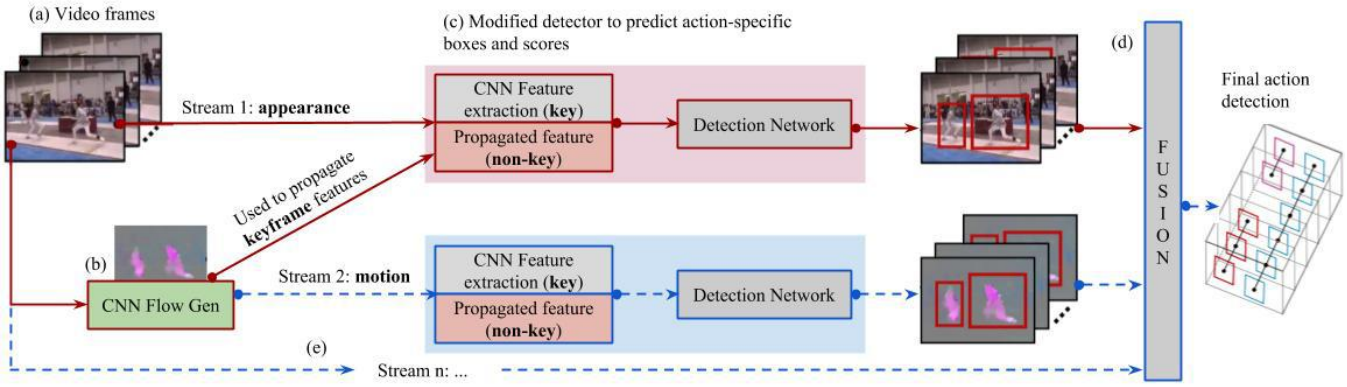


Figure 1: Illustration of the proposed multi-stream action detection pipeline. This architecture takes consecutive video frames as input (**a**). The motion stream is estimated using the flow CNN, which is trained with other network components (**b**). Both the appearance and motion streams are fed to their respective action detectors to predict class-specific action scores and detection boxes (**c**). A chosen fusion technique is used to merge detection results from different streams (**d**). Additional cues can be introduced to the multi-stream framework in a similar way to enhance detection (**e**). The concept of propagated features is detailed in Section 3.2. In the pipeline, **red solid arrows** correspond to tasks already developed and evaluated in this paper.

3.2. Sparse feature propagation

In videos, image content varies slowly over consecutive frames. The similarity is even higher in the CNN deep feature maps which encode high level semantics [16]. It would be costly to apply the full CNN feature extraction and action detection for every video frame exhibiting high data redundancy. As a starting point toward efficient action detection, we exploit neighboring frames' coherence to reduce computation, as was applied by Zhu et al. [16] for video object detection.

As illustrated in Fig. 2, during inference the deeper and more expensive feature extraction network only runs on sparse key frames. Instead of being extracted from the feature network, the feature maps of non-keyframes are propagated from their preceding keyframes. This is achieved by the manner of spatial warping for all locations and channels in the feature maps using optical flows. The flow fields are also estimated by a CNN, trained along with the feature extraction and detection network. Computation reduction can be achieved as CNN flow estimation is fast and inexpensive compared to CNN deep feature extraction.

4. Experiments

We conducted two experiments: action detection for every frame and action detection with sparse feature propagation. At the moment, we **only** qualitatively evaluate the **appearance** stream in both experiments and make observations for future reference. Thorough evaluations will be given to the complete pipeline in the future.

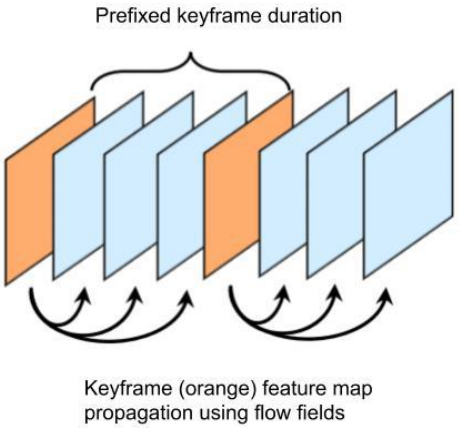


Figure 2: Illustration of flow-propagated feature maps used in our action detection pipeline.

4.1. Dataset

We evaluate action detection on the UCF-101-24 [27] benchmark. It is a subset of UCF-101 which is composed of realistic action videos across 101 action classes from YouTube. The UCF-101-24 consists of 24 classes in 3207 videos which provides frame-level localization annotations. We follow the work of Saha et al. [8], using 2290 of these videos for training and the remaining ones for testing.

4.2. Stand-alone action detection

Our implementation is based on the released code [16] originally used for video object detection. Here, we use

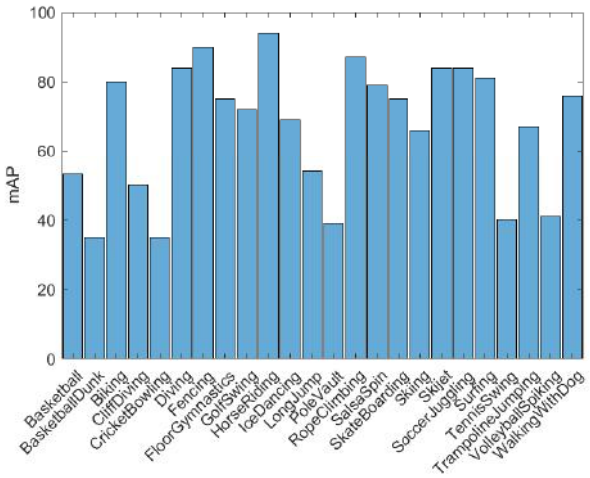


Figure 3: Stand-alone action detection: detection results (mAP) of individual action classes.

ResNet-101 [28] and R-FCN models for convolution feature extraction and action detection respectively. We adopt the training scheme from the original paper. In particular, to train the stand-alone R-FCN in the first experiment, from each training video we uniformly select around 15 frames with annotations. We use ResNet with ImageNet pre-training. During training, we resize all frames to size 600×800 (originally 240×320). For the moment we keep the same set of training parameters used by [16].

Our trained, stand-alone R-FCN achieves a mean Average Precision (mAP) of 67 % (Intersection over Union = 0.5) on average of all actions. The performance of individual action class is reported in Fig. 3. In particular, we observe that certain classes perform significantly worse than others. This can be caused by the low-quality frames and ambiguous context when only learning from the appearance stream. For example, the action Basketball was often misclassified as TennisSwing possibly due to the similarity of the playground (Fig. 4a). Similarly, action RopeClimbing is initially classified as FloorGymnastics until the emergence of a clear rope (Fig. 4b). On the other hand, the action Fencing consistently perform well due to having unique appearances (i.e., white gears) that would not be easily confused with other action classes. We hypothesize that the inclusion of the motion stream will lead to improved classification.

4.3. Action detection with sparse feature propagation

In the second experiment, we apply ResNet-101, R-FCN and FlowNet models for convolution feature extraction, action detection and flow estimation respectively. The entire network is trained end-to-end. In each mini-batch, a pair of nearby video frames (I_r and I_i) are randomly sampled,

(a) Incorrect detection

(b) Incorrect detection

(c) Correct detection

Figure 4: Stand-alone R-FCN action detection results on UCF-101-24 dataset. (a) both images exhibit similar context, but the correct classification of the left image is "Basketball". Likewise in (b), both images share similar context, but the correct action for the left image is "RopeClimbing".

one being the reference frame. The deep feature map f_r is obtained from the reference frame, while FlowNet runs on both frames to estimate the flow field. The estimated flow is then used to propagate f_r to f_i, which is the feature map fed to the detection network. The incurred localization and classification loss are then back-propagated to update all components of the detection and flow estimation network. Here, we use ResNet with ImageNet pre-training. FlowNet is pre-trained on the Flying Chair dataset [17].

In Fig. 5 we show some action localization results when considering data redundancy between nearby frames. In this experiment, every 10 frame is sampled as a keyframe. When tested on a NVIDIA Titan X GPU, each frame detection took **0.022** seconds on average. Our results demonstrate an average of 3 times speedup compared to the stand-alone R-FCN (around **0.057** seconds per frame) while achieving a mAP of 62 % .

5. Future work

In this paper we introduce our motivation and pipeline to address video action detection. Several important aspects are left for further development and exploration.

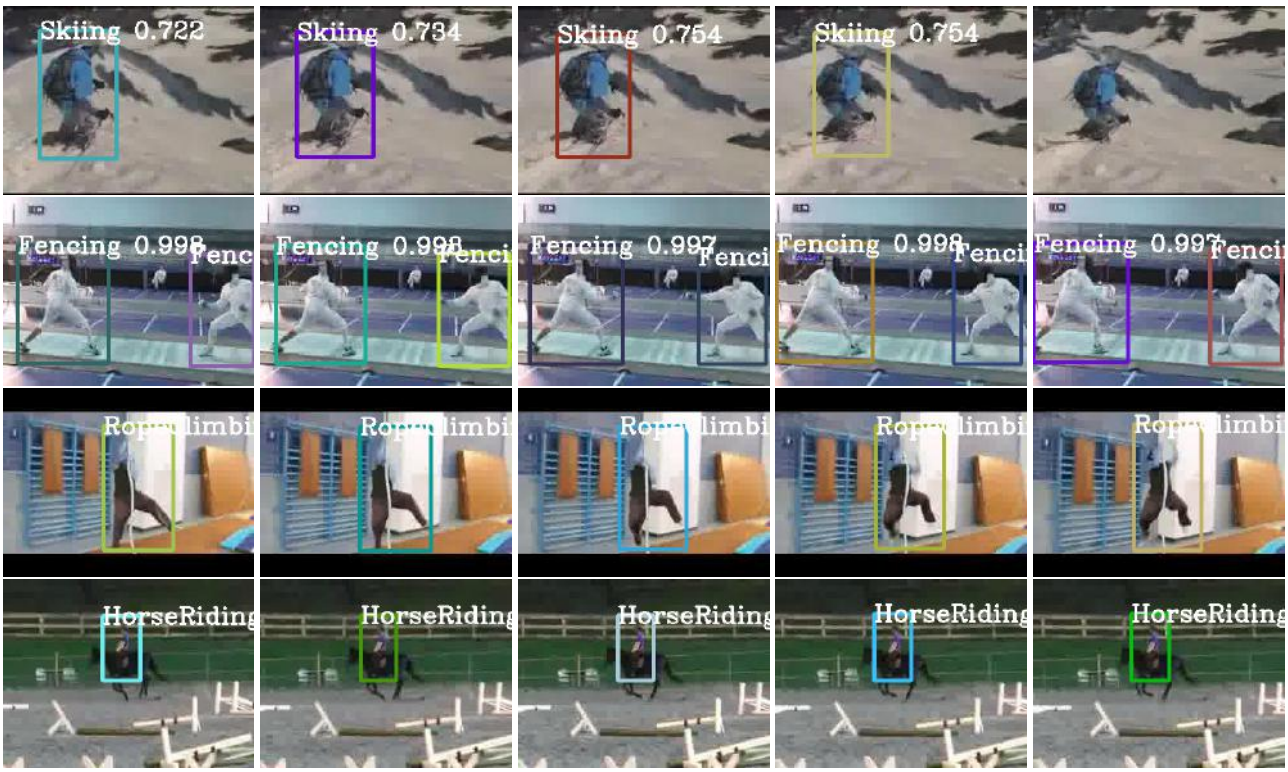


Figure 5: Action detection results on UCF-101-24 dataset. The first column corresponds to detection results on the keyframes. The other four columns correspond to results of the following frames obtained with the propagated features.

We will first quantitatively validate the effect of alleviating data redundancy on action datasets, as temporal continuity may be more crucial in action videos than object videos. Moreover, our method, in terms of computational efficiency, may further benefit from smarter region proposal algorithms that target human presence in early frames. In addition, we will adapt our framework into a multi-stream framework starting with appearance and motion. We will also explore the use of more modalities (e.g., pose) and feature aggregation across frames to capture more temporal information.

Finally, we will perform further computation optimization such as using a smaller base network, one-shot action detector, and optimized hardware implementation for the final smart camera deployment. We believe incorporating the above tasks in hand will lead to a robust and efficient action detection solution suitable for embedded devices.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[2] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, Springer, 2016, pp. 21–37.

[5] S. Saha, G. Singh, M. Sapienza, P. H. S. Torr, and F. Cuzzolin, "Deep learning for detecting multiple space-time action tubes in videos," in *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016.

[6] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, "Action tubelet detector for spatio-temporal action localization," *ICCV, Oct*, vol. 2, 2017.

[7] X. Peng and C. Schmid, "Multi-region two-stream r-cnn for action detection," in *European Conference on Computer Vision*, Springer, 2016, pp. 744–759.

[8] G. Singh, S. Saha, M. Sapienza, P. Torr, and F. Cuzzolin, "Online real-time multiple spatiotemporal action localisation and prediction," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2017, pp. 3657–3666.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European conference on computer vision*, Springer, 2014, pp. 346–361.

[11] R. Girshick, *Fast r-cnn. ieee international conference on computer vision*, 2015.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[13] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang, "Seq-nms for video object detection," *arXiv preprint arXiv:1602.08465*, 2016.

[14] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, *et al.*, "T-cnn: Tubelets with convolutional neural networks for object detection from videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2896–2907, 2018.

[15] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3038–3046.

[16] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *CVPR*, vol. 1, 2017, p. 3.

[17] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2758–2766.

[18] M. Liu and M. Zhu, "Mobile video object detection with temporally-aware feature maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5686–5695.

[19] G. Gkioxari and J. Malik, "Finding action tubes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 759–768.

[20] T. Kroeger, R. Timofte, D. Dai, and L. Van Gool, "Fast optical flow using dense inverse search," in *European Conference on Computer Vision*, Springer, 2016, pp. 471–488.

[21] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, IEEE, 2017, pp. 2923–2932.

[22] Y. Wang, J. Song, L. Wang, L. Van Gool, and O. Hilliges, "Two-stream sr-cnns for action recognition in videos.," in *BMVC*, 2016.

[23] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[24] Y. Zhu, Z. Lan, S. Newsam, and A. G. Hauptmann, "Hidden two-stream convolutional networks for action recognition," *arXiv preprint arXiv:1704.00389*, 2017.

[25] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 3, 2017.

[26] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "Actionvlad: Learning spatio-temporal aggregation for action classification," in *CVPR*, vol. 2, 2017, p. 3.

[27] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

Learning Scene Geometry for Visual Localization in Challenging Conditions

Nathan Piasco[1,2], Désiré Sidibé[1], Valérie Gouet-Brunet[2] and Cédric Demonceaux[1]

***Abstract*— We propose a new approach for outdoor large scale image based localization that can deal with challenging scenarios like cross-season, cross-weather, day/night and long-term localization. The key component of our method is a new learned global image descriptor, that can effectively benefit from scene geometry information during training. At test time, our system is capable of inferring the depth map related to the query image and use it to increase localization accuracy.**

We are able to increase recall@1 performances by 2.15% on cross-weather and long-term localization scenario and by 4.24% points on a challenging winter/summer localization sequence versus state-of-the-art methods. Our method can also use weakly annotated data to localize night images across a reference dataset of daytime images.

I. Introduction

Visual-Based Localization (VBL) is a central topic in robotics and computer vision applications [1]. It consists in retrieving the location of a visual query according to a known absolute reference. VBL is used in many applications such as autonomous driving, augmented reality, robot navigation or SLAM loop closing. In this paper, we address VBL as an image retrieval problem where an input image is compared to a reference pool of localized images. This image-retrieval-like problem is two-step: descriptor computation for both the query and the reference images and similarity association across the descriptors. Since the reference images are associated to a location, by ranking images according to their similarity scores we obtain an approximate location for the query. Numerous works have introduced image descriptors well suited for image retrieval for localization [2], [3], [4], [5], [6].

One of the main challenges of image-based localization remains the mapping of images acquired under changing conditions: cross-season images matching [7], long-term localization [8], day to night place recognition [9], etc. Recent approaches use complementary information in order to address these visually challenging localization scenarios (geometric information through point cloud [10], [11] or depth maps [12], semantic information [13], [12], [7]). However geometric or semantic information are not always available, especially in robotic applications when the sensor or the computational load on the robot is limited.

In this paper, we propose a image descriptor that learns, from an image, the corresponding scene geometry, in order to deal with challenging outdoor large-scale image-based localization scenarios. We introduce geometric information during the training step to make our new descriptor robust to visual changes that occur between images taken at different times. Once trained, our system is only used on images to construct a expressive descriptor for image retrieval. This kind of system design is also known as side information learning [14], as it uses geometric and radiometric information only during the training step and just radiometric data for the image localization. Our method is especially well-suited for robotic long-term localization when the perceptive sensor on the robot is limited to a camera [15], while having access to the full scene geometry off-line [16], [17], [18].

The paper is organized as follows. In section II, we first revisit recent works related to our method, including: state of the art image descriptors for large scale outdoor localization, method for localization in changing environment and side information learning approaches. In section III, we describe in detail our new image descriptor trained with side depth information. We illustrate the effectiveness of the proposed method on four challenging scenarios in section IV. Section V finally concludes the paper.

II. Related Work

Image descriptor for outdoor visual localization. Standard image descriptors for image retrieval in the context of image localization are usually built by combining sparse features with an aggregation method, such as BoW or VLAD. Specific features re-weighting scheme dedicated to image localization have been introduced in [19]. Authors of [20] introduce a re-ranking routine to improve the localization performances on large-scale outdoor area. More recently, [2] introduces NetVLAD, a convolutional neural network that is trained to learn a well-suited image representation for image localization. Numerous other CNN image descriptors have been proposed in the literature [3], [4], [5], [21], [6] and achieve state of the art results in image retrieval for localization. Therefore we use CNN image descriptors as base component in our system.

Localization in challenging condition. In order to deal with visual changes in images taken at different times, [22] uses a combination of handcrafted and learned descriptors. [23] introduces temporal consistency by using a sequence of images, while in our proposal we use only one image as input for our descriptor. In [24], authors synthesize new images to match the appearance of reference images, for instance they synthesized daytime images from night images. Numerous works [25], [8], [7] enhance their visual descriptors by adding semantic information. Although semantic represen-

[1] Le2i, ERL CNRS VIBOT 6000, Université Bourgogne Franche-Comté
[2] Univ. Paris-Est, LaSTIG MATIS, IGN, ENSG, F-94160 Saint-Mandé, France

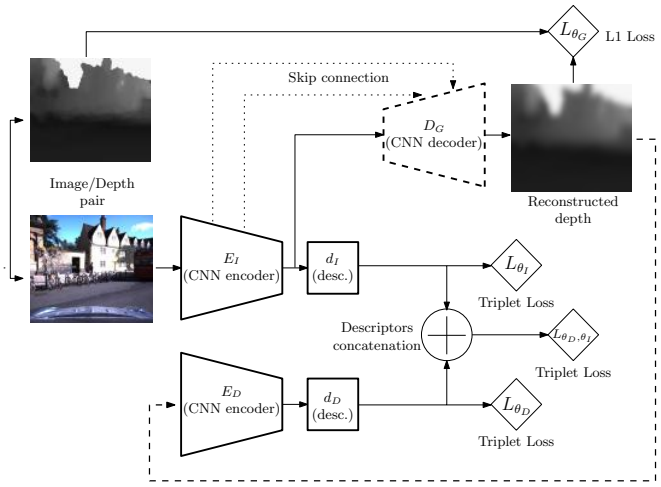


Fig. 1. **Image descriptors training with auxiliary depth data (our work):** two encoders are used for extracting deep features map from the main image modality and the auxiliary reconstructed depth map (inferred from our deep decoder). These features are used to create intermediate descriptors that are finally concatenated in one final image descriptor.

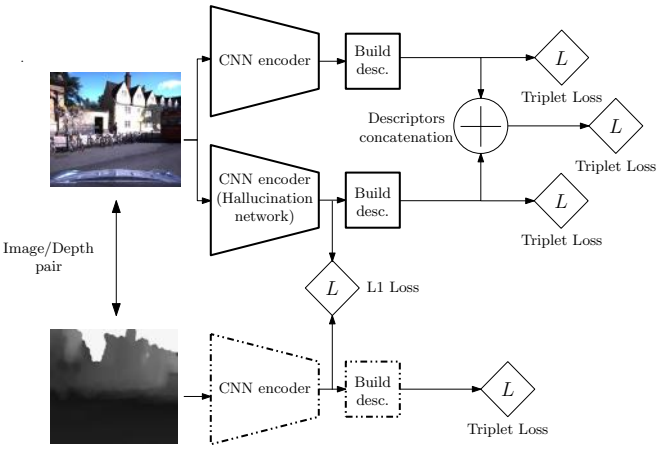


Fig. 2. **Hallucination network for image descriptors learning:** we train an hallucination network, inspired from [14], for the task of global image description. Unlike the proposed method (see figure 1), hallucination network reproduces feature maps that would have been obtained by a network trained with depth map rather than the deep map itself.

tation is robust for long term localization, it may be costly to obtain. Other methods rely on geometric information like point clouds [10], [11], or 3D structures [9]. Geometric information has the advantage of remaining more stable across time comparing to visual information but is not always available. That is why we decide to use depth information as side information in combination with radiometric data to learn a powerful image descriptor.

Learning with side information. Recent work from [26] casts the side information learning problem as a domain adaptation problem, where source domain includes multiples modalities and the target domain is composed of a single modality. Another successful method have been introduced in [14]: authors train a deep neural network to hallucinate features from a depth map only presented during the training process to improve objects detection in images. The closest work to ours, presented in [27], uses recreated thermal images to improve pedestrian detection on standard images only. Our system, inspired by [27], learns how to produce depth maps from images to enhance the description of these images.

III. Method

A. Overview

We design a new global image description for the task of image-based localization. We first extract dense feature maps from an input image with a convolutional neural network encoder (E_I). These feature maps are subsequently used to build a compact representation of the scene (d_I). State-of-the-art features aggregation methods can be used to construct the image descriptor, such as MAC [5] or NetVLAD [2]. We enhance this standard image descriptor with side depth map information that is only available during the training process. To do so, a deep fully convolutional neural network decoder (D_G) is used to reconstruct the corresponding depth map according to the input image. The reconstructed depth is then used to extract a global depth map descriptor. We follow the same procedure used before: we extract deep feature maps with an encoder (E_D) before building the descriptor (d_D). Finally, the image descriptor and the depth map descriptor are L_2 normalized to be concatenated into a single global descriptor. Figure 1 summarizes the whole process of our method. Once trained with geometric and radiometric information, the proposed method is used on images only, to create a descriptor tuned for image localization.

B. Training routine

Trainable parameters are θ_I the weights of encoder and descriptor $\{E_I, d_I\}$, θ_D the weights of the encoder and descriptor $\{E_D, d_D\}$ and θ_G the weights of the decoder used for depth map generation.

For training our system, we follow standard procedure of descriptor learning based on triplet margin losses [2]. A triplet $\{q_{im}, q_{im}^+, q_{im}^-\}$ is composed of an anchor image q_{im}, a positive example q_{im}^+ representing the same scene as the anchor and an unrelated negative example q_{im}^-. The first triplet loss acting on $\{E_I, d_I\}$ is:

$$L_{\theta_I} = max\left(0, \lambda + \left\|f_{\theta_I}(q_{im}) - f_{\theta_I}(q_{im}^+)\right\|_2 - \left\|f_{\theta_I}(q_{im}) - f_{\theta_I}(q_{im}^-)\right\|_2\right), \quad (1)$$

where $f_{\theta_I}(x_{im})$ is the global descriptor of image x_{im} and λ an hyper-parameter controlling the margin between positive and negative examples. f_{θ_I} can be written as:

$$f_{\theta_I}(x_{im}) = d_I(E_I(x_{im})), \quad (2)$$

where $E_I(x_{im})$ represents the deep feature maps extracted by the decoder and d_I the function used to build the final descriptor from the feature.

We train the depth map encoder and descriptor $\{E_D, d_D\}$ in a same manner, equation (1) becoming:

$$L_{\theta_D} = max\left(0, \lambda + \left\| f_{\theta_D}(\hat{q}_{depth}) - f_{\theta_D}(\hat{q}^+_{depth}) \right\|_2 - \left\| f_{\theta_D}(\hat{q}_{depth}) - f_{\theta_D}(\hat{q}^-_{depth}) \right\|_2\right), \quad (3)$$

where $f_{\theta_D}(x_{depth})$ is the global descriptor of depth map x_{depth} and $\hat{x}_{depth}$ is the reconstructed depth map of image x_{im} by the decoder D_G:

$$\hat{x}_{depth} = D_G(E_I(x_{im})). \quad (4)$$

Decoder D_G uses the deep representation of image x_{im} computed by encoder E_I in order to reconstruct the scene geometry. Notice that even if the encoder E_I is not especially trained for depth map reconstruction, its intern representation is rich enough to be used by the decoder D_G for the task of depth map inference. We choose to use the features already computed by the first encoder E_I instead of introducing another encoder for saving computational resources.

The final image descriptor is trained with the following loss:

$$L_{\theta_I,\theta_D} = max\left(0, \lambda + \left\| F_{\theta_I,\theta_D}(q_{im}) - F_{\theta_I,\theta_D}(q^+_{im}) \right\|_2 - \left\| F_{\theta_I,\theta_D}(q_{im}) - F_{\theta_I,\theta_D}(q^-_{im}) \right\|_2\right), \quad (5)$$

where $F_{\theta_I,\theta_D}(x_{im})$ denotes the concatenation of image descriptor and depth map descriptor:

$$F_{\theta_I,\theta_D}(x_{im}) = \left[f_{\theta_I}(x_{im}), f_{\theta_D}(\hat{x}_{depth})\right]. \quad (6)$$

In order to train the depth map generator, we use a simple L_1 loss function:

$$L_{\theta_G} = \left\| x_{depth} - \hat{x}_{depth} \right\|_1. \quad (7)$$

The whole system is trained according to the following constraints:

$$(\theta_I, \theta_D) := arg\min_{\theta_I,\theta_D} \left[L_{\theta_I} + L_{\theta_D} + L_{\theta_I,\theta_D}\right], \quad (8)$$

$$(\theta_G) := arg\min_{\theta_G} \left[L_{\theta_G}\right]. \quad (9)$$

We use two different optimizers: one updating θ_I and θ_D weights regarding constraint (8) and the other updating θ_G weights regarding constraint (9). Because decoder D_G relies on feature maps computed by encoder E_I (see equation (4)), at each optimization step on θ_I we need to update decoder weights θ_G to take in account possible changes in the image features. We finally train our entire system, by alternating between the optimization of weights $\{\theta_I, \theta_D\}$ and $\{\theta_G\}$ until convergence.

C. Hallucination network for image description

We compare our method of side information learning with a state-of-the-art approach system, named hallucination network [14]. The hallucination network is originally designed for object detection and classification in images. We adapt the work of [14] to create an image descriptor system that benefits from depth map side modality during training.

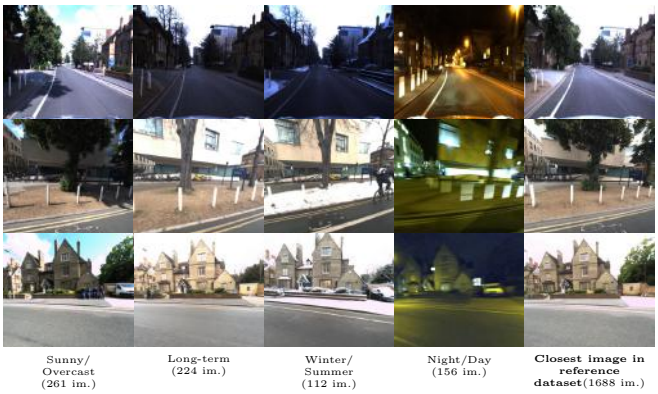


Fig. 3. **Examples of test images :** we evaluate our proposal on four challenging localization sequences. The number under the query set name indicates the amount of query images to compare against the 1688 reference images.

Like our proposal, the trained hallucination network is used on images only and produce a global descriptor for image localization. The system is presented in figure 2. The main difference with our proposal is that the hallucination network reproduces feature maps that would have been obtained by a network trained with depth map rather than the deep map itself. We refer readers to [14] for more information about the hallucination network.

D. Advantages and drawbacks

One advantage of the hallucination network over our proposal is that it does not require a decoder network, resulting on a architecture lighter than ours. However, it needs a pre-training step, where image encoder and depth map encoder are trained separately from each other before a final optimization step with the hallucination part of the system. Our system do not need such initialization. Training the hallucination network requires more complex data than the proposed method. Indeed, it needs to gather triplets of image, and depth map pairs, which require to know the absolute position of the data [2], [6], or to use costly algorithms like Structure from Motion (SfM) [28], [5], [3].

One advantage of our method over the hallucination approach is that we have two unrelated objectives during training: learning a efficient image representation for localization and learning how to reconstruct scene geometry from an image. It means we can train several parts of our system separately, with different source of data. Especially, we can improve the scene geometry reconstruction task with non localized {*image, depth map*} pairs. These weakly annotated data are easier to gather than triplet, as we only need calibrated system capable of sensing radiometric and geometric modalities at the same time. We will show in practice how this can be exploited to fine tune the decoder part to deal with complex localization scenarios in part IV-C.

IV. Experiments

A. Dataset

We have tested our new method on the *Oxford Robotcar* public dataset [17]. This is a common dataset used for image-

based localization [10] and loop closure algorithm involving neural networks training [24].

Training data. We use the temporal redundancy present in the dataset to build the images triplets to train our CNN. We build 400 triplets using three runs acquired at dates: `2015-05-19`, `2015-08-28` and `2015-11-10`. We selected an area of the city different from the one used for training our networks for validation. Depth modality is extracted from the lidar point cloud dataset of *Oxford Robotcar*. When re-projected in the image frame coordinate, it produces a sparse depth map. Since deep convolutional neural networks require dense data as input, we pre-process these sparse modality maps with inpainting algorithm from [29] in order to make them dense.

Testing data. We propose four testing scenarios on the same spatial area (different from the area used for training and validation). The reference dataset is composed of 1688 images taken every 5 meters along a path of 2 km, when the weather was overcast. The four query sets are:

- Sunny/Overcast queries have been acquired during a sunny day.
- Long-term queries have been acquired 7 months after the reference images under similar weather conditions.
- Winter/Summer queries have been acquired during a snowy day.
- Night/Day queries have been acquired at night, resulting in radical visual changes compared to the reference images.

Query examples are presented in figure 3.

Evaluation metric. For a given query, the reference images are ranked according to the cosine similarity score computed over their descriptors. To evaluate the localization performances, we consider two evaluation metrics:

a) Recall @N: we plot the percentage of well localized queries regarding the number N of returned candidates. A query is considered well localized if one of the top N retrieved images lies inside the $25m$ radius of the ground truth query position.

b) Top-1 recall @D: We compute the distance between the top ranked returned database image position and the query ground truth position, and report the percentage of queries located under a threshold D (from 15 to 50 meters), like in [30]. This metric qualifies the accuracy of the localization system.

B. Implementation details

Our proposal is implemented by using Pytorch as deep learning framework, ADAM stochastic gradient descent algorithm for the CNN training with learning rate set to 1e-4, weight decay to 1e-3 and λ in triplet loss equations (1), (3), (5) equal to 0.1. We use batch size between 10 and 25 triplets depending of the size of the system to train, convergence occurs rapidly and takes around 30 to 50 epochs. We perform both positive and negative hard mining, as in [5]. Images and depth maps are re-sized to 224×224 pixels before training and testing.

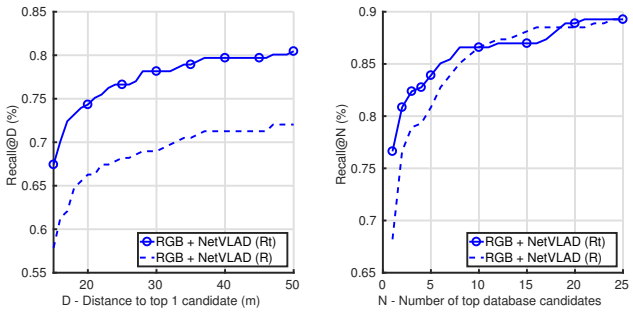


Fig. 4. **Resnet18 (R) versus truncated Resnet18 (Rt) in combination with NetVLAD pooling:** we show the importance of the spatial resolution of the deep feature maps of the encoder used with NetVLAD layer. The truncated version of Resnet18, more than two times lighter the complete one, achieves much better localization results.

Encoder architectures. We test the fully convolutional part of Alexnet and Resnet18 architectures for features extraction. The size of the final features block is $256 \times 13 \times 13$ for Alexnet and $512 \times 7 \times 7$ for Resnet. Initial weights are the ones obtained by training the whole network on ImageNet dataset. We always use Alexnet encoder to extract features from raw depth map, reconstructed depth map, or hallucinated depth map. Indeed the quality of our depth map is usually very low, we have found that using deeper network does not significantly improve localization results.

Descriptor architectures. We test the two state-of-the-art image descriptors MAC [5] and NetVLAD [2]. MAC is a simple global pooling method that takes the maximum of each feature map from the encoder output. NetVLAD is a trainable pooling layer that mimics VLAD aggregation method. For all the experiments, we set the number of NetVLAD clusters to 64. Finally, both MAC and NetVLAD descriptors are L_2 normalized.

Decoder architecture. The decoder used in our proposal is based on Unet architecture and inspired by network generator from [31]. Dimension up-sampling is performed through inverse-convolutions layers. Decoder weights are initialized randomly.

C. Results

Baselines. We compare our method with two state-of-the-art baselines:

*a) RGB only (**RGB**):* simple networks composed of encoder + descriptor trained with only images, without side depth maps information. We evaluate 4 variants of networks, by combining Alexnet (A) or Resnet (R) encoder with MAC or NetVLAD descriptor pooling.

*b) RGB with Depth side information (**RGBtD**):* networks that use pairs of aligned image and depth map during training step and images only at test time. We compare our proposal with our version of hallucination network [14] (hall). We follow training procedure of [14] to train the hallucination network, whereas our proposal is trained as explained in III-B.

Truncated Resnet. We experimented that NetVLAD

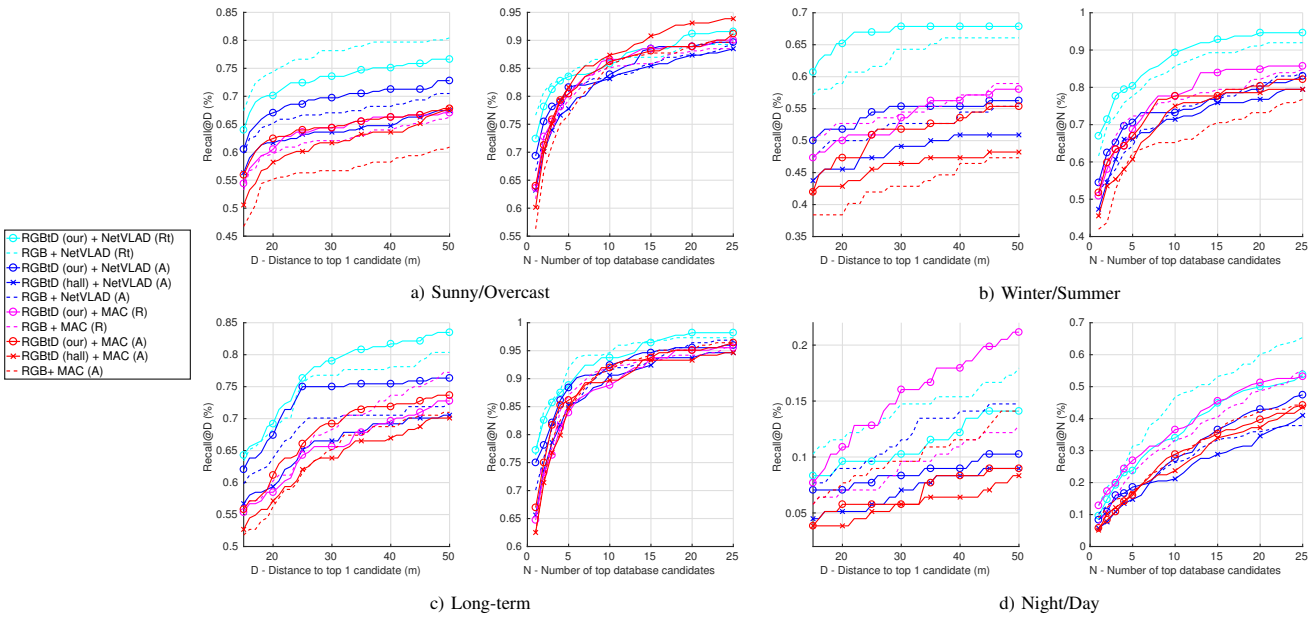


Fig. 5. **Comparison of our method versus hallucination network and networks trained with only images**: our method (-o-) is superior in almost every scenario facing hallucination network (-x-). It also beats, with a significant margin, networks trained with only images (−−). NetVLAD descriptors (blue and cyan curves) are superior to MAC (red and magenta curves), specially in terms of accuracy (Recall@D curve). Night/day dataset remains the most challenging one. Curves best viewed in colors.

descriptor combined with Resnet architecture, RGB + NetVLAD (R), does not perform well. NetVLAD can be view as a pooling method that acts on local deep features densely extracted from the input image. We argue that the spatial resolution of the features block obtained with Resnet encoder is too low compared to the other architecture (for instance 13×13 for Alexnet compared to 7×7 for Resnet for an 224×224 input image). We propose a truncated version of Resnet encoder (Rt), created by drooping the end of the network after the 13th convolutional layer. Thus we obtain a feature block with greater spatial resolution: $256 \times 14 \times 14$ compared to $512\times7\times7$. Recall results on the *Sunny/Overcast* query set for both architectures are presented in figure 4. As the truncated version of Resnet encoder clearly dominates the full one, we use the truncated version for the following experiments.

Discussion. Localization results on the four query sets are presented in figure 5. When we use encoder architecture Alexnet combined with MAC descriptor, our system obtains a mean of +5% on recall@1 score comparing to hallucination (on sun, long-term and winter query sets). The proposed method is also more accurate, obtaining a mean of +2.3% on top-1 recall@15 and +3.7% on top-1 recall@50 scores. These results demonstrate the superiority of our method compared to hallucination for the task of image localization. Both of the RGBtD approaches increase localization results facing the RGB baseline (+7.3% recall@1, +5.6% top-1 recall@15, +5.9% top-1 recall@50 for RGBtD (ours) versus RGB).

Results show that NetVLAD pooling is much more effective than MAC. However, we have not been able to obtain convincing results for hallucination network with Alexnet encoder combined with NetVLAD pooling: reported results show decreasing performances compared to RGB + NetVLAD (A) baseline. In contrast, our proposal achieves superior localization scores with setting Alexnet encoder with NetVLAD descriptor compared to the RGB baseline: 62.85% / **66.27**% recall@1, 55.64% / **57.53**% top-1 recall@15 and 65.91% / **68.46**% top-1 recall@50 (RGB + NetVLAD (A) mean score / **RGBtD (ours) + NetVLAD (A)** mean score on the sun, long-term and winter query sets).

Best localization results are obtained by combining NetVLAD descriptor with truncated Resnet encoder. With this setting, our system achieves 71.53% / **72.2**% recall@1, 62.65% / **63.0**% top-1 recall@15 and 75.63% / **76.0**% top-1 recall@50 (RGB + NetVLAD (Rt) mean score / **RGBtD (ours) + NetVLAD (Rt)** mean score on the sun, long-term and winter query sets).

Considering all the networks, our method shows the best localization improvement on the Winter/Summer query set (figure 5-b): 51.78% / **56.05**% recall@1, 48.21% / **50.0**% top-1 recall@15, 56.92% / **59.38**% top-1 recall@50 (RGB mean score / **RGBtD (ours)** mean score). Standard image descriptors are confused by local changes caused by the snow present on the scene whereas our descriptor remains confident by reconstructing the geometric structure of the scene. Similar results should be intended regarding Night/Day query sets (figure 5-d), however our proposal is not able to improve localization accuracy for this query set. We investigate the night to day localization scenario in the following.

Night to day localization. Night to day localization is an extremely challenging problem: our best RGB baseline achieves less than 13% recall@1. This can be explained by the huge difference in visual appearance between night

TABLE I

CONTRIBUTION OF THE DEPTH SIDE INFORMATION DURING TRAINING.

Query set	Network		Top-1 recall@D			Recall@N	
	Name	#Param.	@15	@30	@50	@1	@5
Sunny/ Overcast	RGB + MAC	2.5M	46.7	56.7	60.9	56.3	76.6
	RGB$^+$ + MAC	7.9M	51.0	61.0	66.7	60.1	79.3
	RGBtD + MAC	7.9M	**55.9**	**64.4**	**67.8**	**64.0**	**80.5**
Long-term	RGB + MAC	2.5M	51.8	65.2	71.0	62.5	84.4
	RGB$^+$ + MAC	7.9M	54.5	68.3	72.3	**67.0**	82.6
	RGBtD + MAC	7.9M	**55.8**	**69.2**	**73.7**	**67.0**	**86.2**
Winter/ Summer	RGB + MAC	2.5M	38.4	43.0	47.3	42.0	62.5
	RGB$^+$ + MAC	7.9M	36.6	42.0	43.0	41.1	56.3
	RGBtD + MAC	7.9M	**42.0**	**51.8**	**55.4**	**51.8**	**67.0**

and daytime images, as illustrated in figure 3. Our system should be able to improve the RGB baseline relying on the learned scene geometry, which remains the same during day and night. Unfortunately, we use training data exclusively composed of daytime images, thus making the decoder unable to reconstruct a depth map from an image taken at night. The last line of figure 6 shows the poor quality of decoder output after initial training. In order to improve the decoder's performances, we propose to use weakly annotated data to fine tune the decoder part of our system. We collect 1000 pairs of image and depth map acquired at night and retrain only decoder weights θ_G using loss of equation (7). Figure 6 presents the qualitative amelioration on the inferred depth map after the fine tuning. Such post-processing trick cannot be used to improve standard RGB image descriptors, because we need to know the location of the night data. For instance, we use a night run from the Robotcar dataset with a low quality GPS signal, that makes impossible the automatic creation of triplets that are essential for training a deep image descriptor. We show in figure 7 that we are able to nearly multiply by two the localization performances by only fine tuning a small part of our system. Our best network achieves 23% recall@1 against 13% recall@1 for the best RGB baseline.

Contribution of the depth information. In this paragraph, we investigate the impact on localization performances provided by the side geometry information on our method. To ensure a fair comparison in terms of number of trainable parameters, we introduce RGB$^+$ network that has the same architecture as our proposed method. We train RGB$^+$ with images only to compare the localization results against our method that uses side depth information. For training RGB$^+$, we simply remove the loss introduced in equation (4), and make the weights of the decoder trainable when optimizing triplets losses constraints. Results of this experiment with encoder architecture Alexnet are presented in table I.

Increasing the size of the system results in a better localization (RGB$^+$ + MAC versus RGB + MAC) on the two easiest query sets. Surprisingly RGB$^+$ system decreases localization performances on the winter queries compared to RGB. The system has probably over-fitted on the training data that are visually close to queries of "Sunny" set and "Long-term" set, but quiet different from the queries of

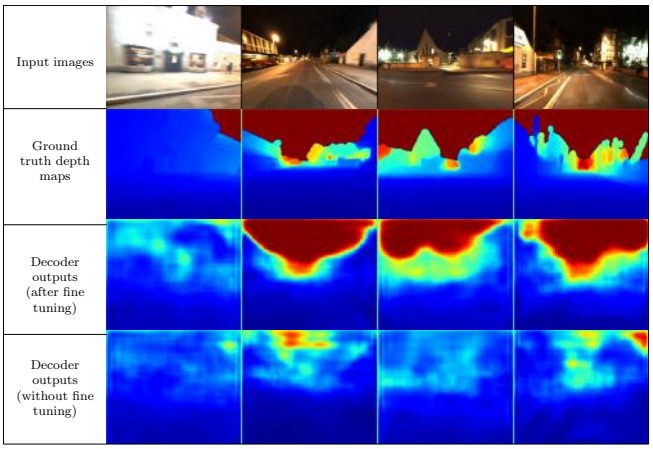


Fig. 6. **Effect of fine tuning with night images on decoder output:**. Decoder trained with daylight images is unable to reconstruct the scene geometry (bottom line). Fine tuning the network with less than 1000 pairs {image, depth map} acquired by night highly improves appearance of the generated depth maps. Maps best viewed in color.

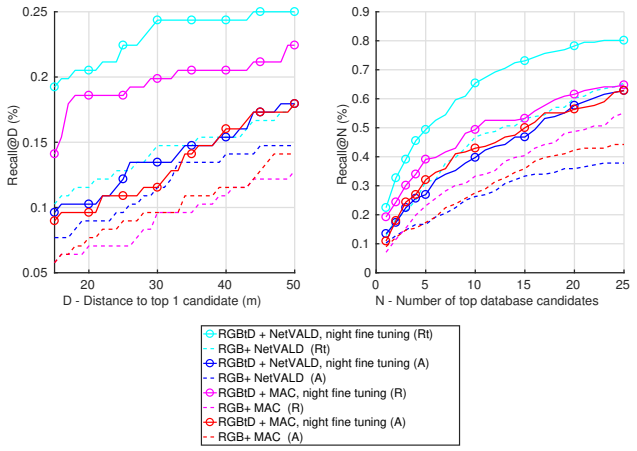


Fig. 7. **Results on Night/Day query set after fine tuning:** we are able to drastically improve localization performance for the Night/Day challenging scenario by only fine tuning the decoder part of our network with weakly annotated data. Curves best viewed in color.

"Winter" set (see figure 3). Our RGBtD + MAC system always produces higher localization results facing RGB$^+$ + MAC, which shows that the side depth information provided during training is wisely used to describe the image location.

V. CONCLUSION

We have introduced a new competitive global image descriptor designed for image-based localization under challenging conditions. Our descriptor handle visual changes between images by learning the geometry of the scene. Strength of our method remains in the fact that it needs geometric information only during the learning procedure. Our trained descriptor is then used on images only. Experiments show that our proposal is much more efficient than state-of-the-art localization methods [2], [5], including methods based on side information learning [14]. Our descriptor performs especially well for challenging cross-season localization scenario, therefore it can be used to solve long-term place recognition

problem. We additionally obtain encouraging results for night to day image retrieval.

In a future work we will investigate the use of other modalities as side information sources, like the reflectance factor provided by lidars. We also want to study the generalization capability of our system, by considering a different image-based localization task like direct pose regression [32].

REFERENCES

[1] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet, "A survey on Visual-Based Localization: On the benefit of heterogeneous data," *Pattern Recognition*, vol. 74, pp. 90–109, feb 2018. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0031320317303448

[2] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 5297–5307, 2017. [Online]. Available: http://arxiv.org/abs/1511.07247

[3] H. J. Kim, E. Dunn, and J.-M. Frahm, "Learned Contextual Feature Reweighting for Image Geo-Localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "End-to-End Learning of Deep Visual Representations for Image Retrieval," *International Journal of Computer Vision (IJCV)*, vol. 124, no. 2, pp. 237–254, 2017.

[5] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN Image Retrieval with No Human Annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. [Online]. Available: http://arxiv.org/abs/1711.02512

[6] L. Liu, H. Li, and Y. Dai, "Deep Stochastic Attraction and Repulsion Embedding for Image Based Localization," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2018. [Online]. Available: https://arxiv.org/pdf/1808.08779.pdf

[7] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware Visual Localization under Challenging Perceptual Conditions," *Proceedings of the IEEE International Conference of Robotics and Automation (ICRA)*, pp. 2614–2620, 2017.

[8] C. Toft, E. Stenborg, L. Hammarstrand, L. Brynte, M. Pollefeys, T. Sattler, and F. Kahl, "Semantic Match Consistency for Long-Term Visual Localization," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2018.

[9] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[10] T. Sattler, W. Maddern, A. Torii, J. Sivic, T. Pajdla, M. Pollefeys, and M. Okutomi, "Benchmarking 6DOF Urban Visual Localization in Changing Conditions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [Online]. Available: http://arxiv.org/abs/1707.09092

[11] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic Visual Localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [Online]. Available: http://arxiv.org/abs/1712.05773

[12] G. Christie, G. Warnell, and K. Kochersberger, "Semantics for UGV Registration in GPS-denied Environments," *arXiv preprint*, 2016. [Online]. Available: http://arxiv.org/abs/1609.04794

[13] S. Ardeshir, A. R. Zamir, A. Torroella, and M. Shah, "GIS-assisted object detection and geospatial localization," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, vol. 8694 LNCS, no. PART 6, 2014, pp. 602–617.

[14] J. Hoffman, S. Gupta, and T. Darrell, "Learning with Side Information through Modality Hallucination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 826–834. [Online]. Available: http://ieeexplore.ieee.org/document/7780465/

[15] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt, "Scalable 6-DOF localization on mobile devices," *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, vol. 8690 LNCS, no. PART 2, pp. 268–283, 2014.

[16] N. Paparoditis, J.-P. Papelard, B. Cannelle, A. Devaux, B. Soheilian, N. David, and E. Houzay, "Stereopolis II: A multi-purpose and multi-sensor 3D mobile mapping system for street visualisation and 3D metrology," *Revue française de photogrammétrie et de télédétection*, vol. 200, no. 1, pp. 69–79, 2012.

[17] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *The International Journal of Robotics Research (IJRR)*, 2016.

[18] S. Wang, M. Bai, G. Mattyus, H. Chu, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler, and R. Urtasun, "TorontoCity: Seeing the World with a Million Eyes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. [Online]. Available: http://arxiv.org/abs/1612.00423

[19] R. Arandjelović and A. Zisserman, "DisLocation : Scalable descriptor," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2014.

[20] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys, "Large-Scale Location Recognition and the Geometric Burstiness Problem," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. J. Milford, "Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free," in *Robotics Science and Systems (RSS)*, 2015.

[22] T. Naseer, W. Burgard, and C. Stachniss, "Robust Visual Localization Across Seasons," *IEEE Transactions on Robotics (TRO)*, vol. 34, no. 2, pp. 289–302, 2018.

[23] S. Garg, N. Suenderhauf, and M. Milford, "Don't Look Back: Robustifying Place Categorization for Viewpoint- and Condition-Invariant Place Recognition," in *Proceedings of the IEEE International Conference of Robotics and Automation (ICRA)*, 2018. [Online]. Available: http://arxiv.org/abs/1801.05078

[24] H. Porav, W. Maddern, and P. Newman, "Adversarial Training for Adverse Conditions: Robust Metric Localisation using Appearance Transfer," in *Proceedings of the IEEE International Conference of Robotics and Automation (ICRA)*, 2018. [Online]. Available: http://arxiv.org/abs/1803.03341

[25] E. Stenborg, C. Toft, and L. Hammarstrand, "Long-term Visual Localization using Semantically Segmented Images," *arXiv*, 2018. [Online]. Available: http://arxiv.org/abs/1801.05269

[26] W. Li, L. Chen, D. Xu, and L. Van Gool, "Visual Recognition in RGB Images and Videos by Learning from RGB-D Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 8, p. 2030 2036, 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8000401/

[27] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised Monocular Depth Estimation with Left-Right Consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [Online]. Available: http://arxiv.org/abs/1609.03677

[29] M. Bevilacqua, J. F. Aujol, P. Biasutti, M. Brédif, and A. Bugeau, "Joint inpainting of depth and reflectance with visibility estimation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 125, pp. 16–32, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.isprsjprs.2017.01.005

[30] A. R. Zamir and M. Shah, "Image geo-localization based on multiplenearest neighbor feature matching using generalized graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36, no. 8, pp. 1546–1558, 2014.

[31] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[32] E. Brachmann and C. Rother, "Learning Less is More - 6D Camera Localization via 3D Surface Regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [Online]. Available: http://arxiv.org/abs/1711.10228

Comparison of region of interest segmentation methods for video-based heart rate measurements

Peixi Li[1], Yannick Benezeth[1], Keisuke Nakamura[2], Randy Gomez[2], Chao Li[3], Fan Yang[1]

[1] *Le2i EA7508, Arts et Métiers, Univ. Bourgogne Franche-Comté, Dijon, France*
[2] *Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako-shi, Saitama, Japan*
[3] *State Key Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences,Beijing 100190, China*

***Abstract*—Conventional contact photoplethysmography (PPG) sensors are not suitable in situations of skin damage or when unconstrained movement is required. As a consequence, remote photoplethysmography (rPPG) has recently emerged because it provides remote physiological measurements without expensive hardware and improves comfort for long-term monitoring. RPPG estimation methods use the spatially averaged RGB values of pixels in a Region Of Interest (ROI) to generate a temporal RGB signal. The selection of ROI is a critical first step to obtain reliable pulse signals and must contain as many skin pixels as possible with a low percentage of non-skin pixels. In this paper, we experimentally compare seven ROI segmentation methods in the perspective of heart rate (HR) measurements with dedicated metrics. The algorithms are compared using our in-house database *UBFC-RPPG*, comprising of 53 videos specifically geared towards rPPG analysis.**

***Keywords*-remote photoplethysmography (rPPG); Heart Rate (HR); Region of Interest (ROI);**

I. INTRODUCTION

THE principle of photoplethysmography (PPG) 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. With the emergence of video-based health care monitoring, remote photoplethysmography (rPPG) has recently been developed [2] as it allows remote physiological measurements only based on the ambient light and a video camera, hence reducing user constraint and without any expensive and specialist hardware requirement. The very low signal-to-noise ratio (SNR), precisely the ratio between the rPPG signal and all possible noises, represents the main difficulty of this methodology. Some research teams have proposed computer vision or signal processing techniques to improve the robustness of the original method [2]. Current rPPG methods are not yet as accurate as the ECG's measurements nevertheless it seems obvious that, in a close future, they will enable a flexible monitoring of human vital signs (e.g. the heart rate, breathing rate) with a significant reduction of user's constraints or eventually extending the duration of the monitoring.

As discussed in different review papers [1], [3], many rPPG methods implement a general pipeline-based framework: 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. This pipeline-based framework emphasizes the importance of the common first step of ROI segmentation. Several approaches have been proposed for ROI selection in the video stream. In earlier studies, manual selection of the ROI have been used [2]. ROI segmentation can also be based on the result of classical face detection [4] and tracking algorithms and possibly refined with skin pixel classification [5].

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 (e.g. in [6]) that the quality of the ROI has a direct impact on the quality of the rPPG signal. 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 equals to the square root of the number of pixels used in the averaging process [7]. Second, the quality is also affected by the percentage of non-skin pixels in the ROI. All rPPG algorithms suffer from performance degradation when the ROI is not properly selected.

From these observations, we propose to quantitatively evaluate several ROI segmentation methods in the perspective of HR measurements with dedicated metrics. The algorithms are compared using the two datasets of our in-house database *UBFC-RPPG* [8], comprising of 53 videos specifically geared towards rPPG analysis.

In section 2, seven commonly-implemented ROI segmentation methods are described. The video dataset used to compare those methods is described in section 3 while results and conclusion are presented in sections 4 and 5.

II. ROI SEGMENTATION ALGORITHMS

Although different, most ROI segmentation techniques are based on the result of classical face detection and tracking algorithms and possibly refined with skin pixel classification or more precise ROI definition based on a set of landmarks. In this section, we present seven implemented ROI segmentation methods.

Face detection (*face*)
The easiest way to segment the ROI is to use directly the detected face. It is important to note that all other methods presented in this paper are based on the face detection result. In this experiment, we have used the well-known Viola-Jones face detector [9]. In order to avoid spurious movements of the detected face, we also use Kanade-Lucas-Tomasi tracking [10] algorithm (*cf.* Fig. 1(a)).

Face cropped (*crop*)
The *face* ROI contains a significant amount of non-skin pixels from the background, the hair or the clothes (cf. Fig. 1(b)). As suggested by [4], it is possible to simply crop the ROI selecting the center 60% width and height of the box as ROI (*cf.* Fig. 1(b)).

Rule-based skin detection (*rule*)
Because in the context of remote heart rate measurements, we are only interested in skin pixels, another obvious refinement of ROI segmentation method is to perform a pixel-based skin classification on the detected face. Many review papers have been published presenting various color-based skin detection rules (*e.g.* in [11]). It is clearly impossible to implement all possible methods so we have selected a fairly common method [12] in this experiment (*cf.* Fig. 1(c)).

$$\begin{aligned} &\text{(R,G,B) is classified as skin if} \\ &R > 95 \text{ and } G > 40 \text{ and } B > 20 \text{ and} \\ &max\{R,G,B\} - min\{R,G,B\} > 15 \text{ and} \\ &|R - G| > 15 \text{ and } R > G \text{ and } R > B \end{aligned} \tag{1}$$

Histogram-thresholding skin detection (*Conaire*)
Macwan et al. [13] uses Conaire et al. [14] in their rPPG algorithm. This skin detector is based on a thresholding of a non-parametric histogram trained using manually annotated skin and non-skin pixels. This method is eventually based on a Look-Up-Table (LUT) and is thus very fast (*cf.* Fig. 1(d)).

Adaptive-range skin detection (adaptive)
Instead of using fixed thresholds for detecting skin pixels, it is possible to derive the range of skin pixel values using specific ROI of the face. For example, in this experiment, a small ROI in the center of the face bounding box is used to create a reference skin color specific to the detected face. Similar pixels in the detected face ROI are then segmented to create the ROI mask. The main advantage of this method is that it is person-specific and does not rely on global threshold values (*cf.* Fig. 1(e)).

Graph-cut based skin segmentation (*Graph-cut*)
All previous methods perform pixel based skin detection. They are usually very fast but are somehow limited because it does not use any spatial information. Another strategy is to formulate the skin segmentation problem as a segmentation problem using for example Graph-cut [15] (*cf.* Fig. 1(f)).

Landmark based skin segmentation (*landmarks*)
Another strategy to select the ROI is to define a polygon based on a set of landmarks. In this experiment, we used the algorithm proposed by Kazemi et al. [16]. We have selected the contour of the face using the detected landmarks as illustrated in Fig. 1(g).

III. VIDEO DATASET

The ROI segmentation algorithms have been evaluated in the framework of remote heart rate measurements. To this end, we have used the two datasets of our in-house database *UBFC-rPPG* [8] comprising of 53 videos. In the first dataset (7 videos), the volunteers were asked to sit still while in the second dataset (46 videos) the subjects were required to play a time sensitive mathematical game that aimed at augmenting their heart rate while simultaneously emulating a normal human-computer interaction scenario. The database which is focused specifically on rPPG analysis was created using a custom C++ application for video acquisition with a Logitech C920 web camera placed at a distance of about 1 meter from the subject. The video was recorded with a frame 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 comprising of the PPG waveform as well as the PPG heart rates. The experimental setup with sample images is depicted in Fig. 2. Video frames synchronized with PPG sensor data can be downloaded from our project page[1].

IV. EXPERIMENTS

A. System framework

For each video frame, the segmented ROI is spatial averaged to obtain the RGB values. The result of this process is a RGB time series. The RGB temporal traces are then pre-processed by zero-mean and unit variance normalization, detrended using smoothness priors approach and band-pass filtered with Butterworth filter. The rPPG signal is then extracted using the chrominance-based method (later called CHROM) [17]. This method applies simple linear combinations of RGB channels and obtains very interesting performance with low computational complexity. Let $y^c(t)$ be the RGB time series obtained after pre-processing, where $c \in \{R, G, B\}$ is the color channel, CHROM method projects RGB values onto two orthogonal chrominance vectors X and Y:

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

[1] https://sites.google.com/view/ybenezeth/ubfcrppg

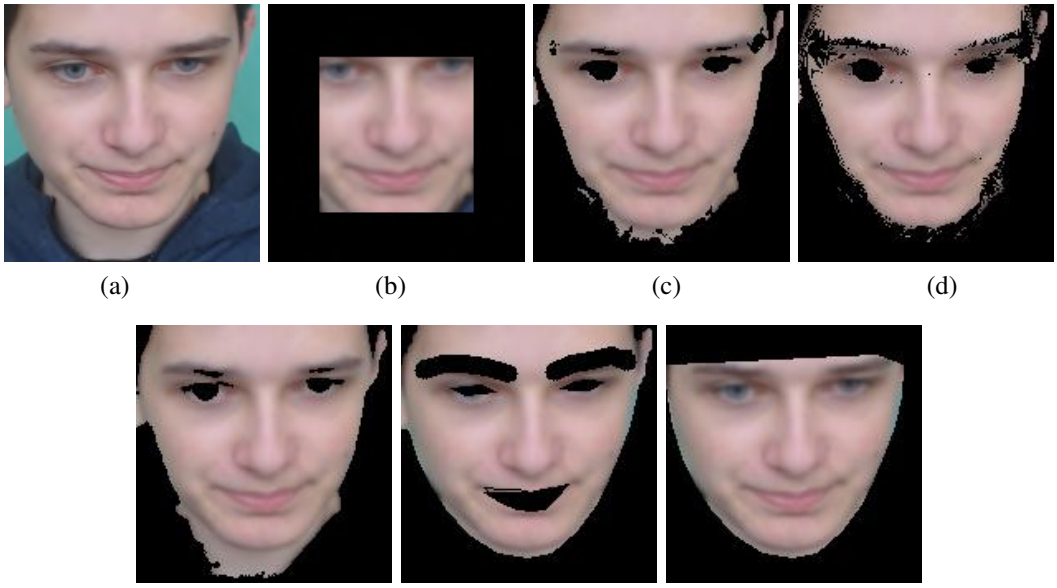


Figure 1. ROI segmentation result examples for (a) face (b) crop (c) rule (d) Conaire (e) adaptive (f) Graph-cut (g) Landmarks.

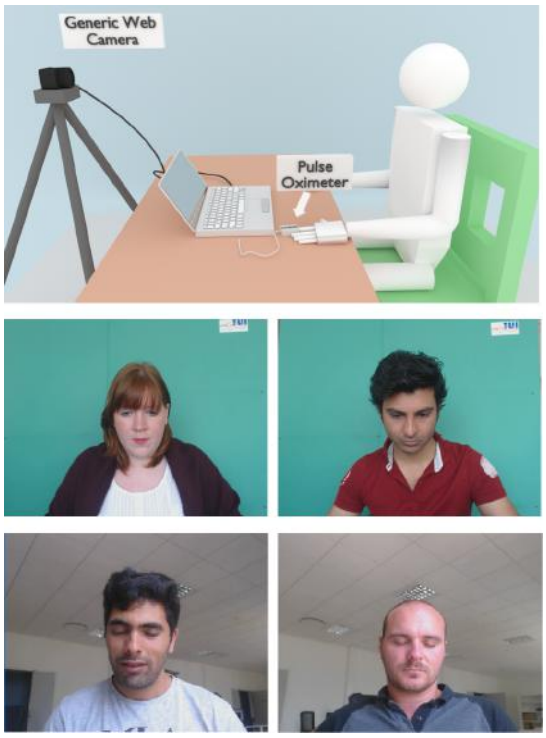


Figure 2. Experimental Setup (top) and sample images from the *UBFC-RPPG* database.

The pulse signal S is finally calculated with $S(t) = X(t) - \alpha Y(t)$ where $\alpha = \sigma(X)/\sigma(Y)$. Because X and Y are two orthogonal chrominance signals, PPG-induced variations will likely be different in X and Y, while motion affects both chrominance signals identically.

For each video, we estimate heart rate in a sliding window framework. Welch's method is used to obtain the periodogram over a 20 seconds 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 and on the rPPG signal given by each evaluated ROI segmentation methods.

B. The Evaluation Metrics

To evaluate the performance of the ROI segmentation method in the perspective of heart rate measurements, the following metrics are used:

- **Mean Absolute Error** (MAE) in beats per minute (bpm) is calculated between heart rate estimated from rPPG signals HR_{rPPG} and heart rate estimated from PPG signals HR_{PPG} with $|HR_{rPPG} - HR_{PPG}|$, averaged per video.
- **Root mean square error** (RMSE) between HR_{rPPG} and HR_{PPG}.
- **Signal-to-Noise Ratio** (SNR) is calculated as the ratio of the power of the main pulsatile component and the power of background noise, computed in dB due to the wide dynamic range.
- **Precision at 5 bpm**. This metric represents the percentage of estimations where the absolute error is under a threshold (2.5 or 5 bpm).

C. Results

Averaged results are presented in Table I. First, it is possible to observe that the ROI selection has indeed a large

impact on the final heart rate measurement precision. One example for this idea is that the SNR obtained with *face* is only 2.11 *dB* while *adaptive* has a SNR of 4.15 *dB*. Second, it is also possible to observe that the overall ranking changes from one metric to another one. *Graph-cut* for example is the best method according to the MAE but performs poorly according to the SNR. From this observation, it is very difficult to give a final ranking of ROI segmentation method. However, we can observe that *face* obtains consistently the worst result. Then, more complex methods such as *Graph-cut* or *landmarks* obtain interesting results but are significantly computationally more expensive than pixel-based methods. *crop* obtains surprisingly good results in this experiment. One possible explanation is that in all videos, the volunteers remained silent and never opened their mouths. The mouth area is very important in the *crop* ROI (*cf.* Figure 1(b)). As a consequence, even if it is not the best for all metrics, the histogram thresholding method proposed by Conaire et al. [14] is an interesting ROI selection. It obtains consistently good performance with all metrics and interestingly, this method is very fast because it is based eventually on a simple LUT.

Table I
THE AVERAGE EVALUATION VALUES FOR ROI DETECTION

	Precision5	**MAE**	**RMSE**	**SNR**
face	0.883	4.43	7.19	2.11
crop	0.938	2.35	3.86	2.98
rule	0.909	3.91	5.34	3.11
Conaire	0.931	3.02	4.99	3.02
adaptive	0.913	3.63	5.50	4.15
Graph-cut	0.941	2.34	3.82	2.95
landmarks	0.922	3.21	4.95	2.87

V. CONCLUSION

ROI segmentation is a critical first step in all remote PPG algorithms to obtain a reliable pulse signal. The ROI must contain as many skin pixels as possible with a low percentage of non-skin pixels. From this observation, we present in this paper a comparative study of seven different ROI selection methods. These methods were implemented and evaluated in the perspective of heart rate measurements. The experiments were done with a low cost camera and a contact PPG as the ground truth. 53 videos were recorded and the averaged results of each method were calculated. The results show that the ROI selection has indeed a large impact on the final heart rate measurement precision and that the histogram thresholding method proposed by Conaire et al. [14] obtains in average very good performance with all metrics. Moreover, this method is interestingly very fast because it is based eventually on a simple LUT.

Even if color-based ROI selection method are fast and reliable, these methods do not consider the distinct pulsatility feature of informative ROI, *i.e.* only skin tissue generates pulsatility. In future work, this particular property may help in improving ROI segmentation methods. Second, it has been shown that the rPPG signal is not distributed homogeneously across the skin, as a consequence the regular spatial averaging of the segmented ROI may not be optimal and offer good opportunities to improve current ROI segmentation techniques.

REFERENCES

[1] Sun, Y. & Thakor, N. Photoplethysmography revisited: from contact to noncontact, from point to imaging. *IEEE Trans. on Biomedical Engineering* **63**, 463–477 (2016).

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

[3] Mcduff, D. J., Estepp, J. R., Piasecki, A. M. & Blackford, E. B. A Survey of Remote Optical Photoplethysmographic Imaging Methods. *int. conf. of the IEEE Engineering in Medicine and Biology Society* .

[4] Poh, M. Z., McDuff, D. J. & Picard, R. W. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Trans. on Biomedical Engineering* (2011).

[5] Wang, W., Stuijk, S. & de Haan, G. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE Trans. on Biomedical Engineering* (2015).

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

[7] Guazzi, A. R., Villarroel, M., Frise, M. C., Robbins, P. a. & Tarassenko, L. Non-contact Measurement of Oxygen Saturation With an RGB Camera. *Journal of Biomedical Optics* **45**, 1764–1771 (2015).

[8] Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A. & Dubois, J. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters* (2017).

[9] Viola, P. & Jones, M. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, I–I (2001).

[10] Lucas, B. D., Kanade, T. *et al.* An iterative image registration technique with an application to stereo vision. *Proceedings DARPA Images Understanding Workshop* 121–130 (1981).

[11] Kakumanu, P., Makrogiannis, S. & Bourbakis, N. A survey of skin-color modeling and detection methods. *Pattern recognition* **40**, 1106–1122 (2007).

[12] Kovac, J., Peer, P. & Solina, F. Human skin color clustering for face detection. In *EUROCON 2003. Computer as a Tool. The IEEE Region 8*, vol. 2, 144–148 (IEEE, 2003).

[13] Macwan, R., Benezeth, Y., Mansouri, A., Nakamura, K. & Gomez, R. Remote photoplethysmography measurement using constrained ica. *IEEE int. conf. on E-Health and Bioengineering* (2017).

[14] Conaire, C. O., O'Connor, N. E. & Smeaton, A. F. Detector adaptation by maximising agreement between independent data sources. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1–6 (2007).

[15] Hu, Z., Wang, G., Lin, X. & Yan, H. Skin segmentation based on graph cuts. *Tsinghua Science & Technology* **14**, 478–486 (2009).

[16] Kazemi, V. & Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1867–1874 (2014).

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

PULSE RATE VARIABILITY FOR EMOTIONAL STATE ASSESSMENT

Rita Meziati Sabour, Yannick Benezeth, Fan Yang

ImViA EA 7535, Univ. Bourgogne Franche-Comté, Dijon, France

ABSTRACT

Both our health condition and psychical state are closely related to the physiological signals that our body sends. Several studies analyzed these physiological signals in order to evaluate their relationship with our emotions. It appeared that these cues are linked to the activity of the autonomic nervous system, which can be particularly well described by the successive accelerations and decelerations in the heart rate. The changing heart rhythms imply variations in the blood volume pulse, which is usually measured with photoplethysmography. Lately, remote photoplethysmography has been introduced and constitutes an affordable and simple pulse sensor. In this study, the blood volume variations that are given by the pulse rate variability are analyzed in order to classify and recognize emotions. We used CAS(ME)^2 dataset, which proposes videos of participants filmed while watching emotion inducing videos. We obtained optimistic results, and an accuracy rate of nearly 77%.

***Index Terms*—remote photoplethysmography, heart rate variability, pulse rate variability, emotion recognition, autonomic nervous system**

I. INTRODUCTION

Our heart rate continuously changes, involving oscillations in the beat-to-beat time intervals. *Heart Rate Variability* (HRV) is the conventional term to describe fluctuations in the heart rate. In the last decades, several studies have demonstrated that sympathovagal activity can be monitored via HRV assessment [14][15]. This is explained by the direct relationship between HRV and the *autonomic nervous system* (ANS), which controls bodily functions such as the heart rate, the respiratory rate, hormonal functions and digestion. ANS is composed of two branches: the *Sympathetic Nervous System* (SNS) and the *Parasympathetic Nervous System* (PNS), which have complementary behaviors. Therefore, since HRV contains information about heart rate acceleration and deceleration, it is an interesting indicator of the ANS activity, which is linked to emotional states [14][17].

The HRV signals are classically measured with electrocardiograms (ECG), which assess the electrical activity of the heart. ECG recordings give consecutive sequences, each containing a QRS complex, and the differences between times of occurrence of successive RR peaks give the HRV signal. The cardiac pumping activity leads to blood volume changes within tissues. *Blood Volume Pulse* (BVP), also called *pulse wave*, can be obtained by using *photoplethysmography* (PPG). PPG sensors optically measure oxygen saturation, and operate either in a transmission or a reflection mode. Typical measurement sites for PPG are areas of the body where transmitted or reflected light can easily be collected, such as fingertips, great toes and earlobes. The amount of absorbed or reflected light depends on changes in the blood volume.

PPG has been extended towards a non-contact use, this is what we call *remote photoplethysmography* (RPPG). This new method consists in measuring the absorption of ambient light from a face video. This is done by analysing subtle color changes in the illuminated skin surface, which are due to blood pulsations caused by the cardiac activity. RPPG has the advantage of being an inexpensive and simple substitute to ECG and PPG. From an RPPG signal, the *pulse rate variability* (PRV) can be estimated [19][20] by calculating the pulse-to-pulse time intervals (PPI).

Both HRV and PRV describe changing heart beat rhythms, and multiple researches focused on the similarity between PRV and HRV [21]. Therefore, since fluctuations in the heart rate are closely linked with changes in the sympathetic-parasympathetic balance, the PRV is an indicator of the activity of the ANS [22]. It is commonly accepted that the high frequency (HF) and the low frequency (LF) components of the HRV, and thus the PRV, reflect the interaction between the sympathetic and the parasympathetic branches of the ANS [22].

Emotion classification has been a challenge for computer vision scientists within the last decades, and proposed methods included gesture tracking, and facial expression analysis [1]. In a parallel way, a great number of computer science researches explored the possibility of classifying emotions based on physiological cues that our body produces. A multimodal approach for emotion recognition fuses EEG, peripheral, electromyogram (EMG), electrooculogram (EOG) and BVP signals, respiration rate and skin temperature in [3]. Other studies focused on the physiological signals that have a direct link to the ANS, including the heart rate [4]. Furthermore, HRV showed its efficiency as an indicator of the autonomic activity in revealing basal stress [6], and even to classify different stress scenarios [7].

The current study proposes to capitalize on the reliability of the PRV to describe changes in the ANS (via the LF and HF component analysis) for emotional state classification and recognition. Experimental computations were applied to CAS(ME)^2 dataset which includes videos of subjects filmed while watching emotion excitation videos. This dataset lists macro and micro-expressions shown by each participant, and is originally aimed at macro and micro-expression spotting and recognition applications. These expressions occur punctually (around $1/4$s for micro and up to 2s for macro expressions), while the PRV describes changes for a longer duration, necessitating the entire videos to extract meaningful information.

The paper is built as follows: PRV signal estimation and feature extraction is detailed in section II. A presentation of CAS(ME)^2 as well as the obtained classification results are explained in section III. The results are then discussed in the same section. Section IV gives a conclusion and perspectives for future works.

II. PRV-based feature extraction

From a video input, three main steps are followed: RGB trace extractraction, pulse wave estimation and PRV signal extraction; as shown in Fig. 1. The PRV high and low frequency components are then computed as explained in this section.

Fig. 1. Flowchart of the PRV-based feature extraction

The face represents our region of interest (ROI). It is located in each video frame using the Viola-Jones algorithm for rapid object detection [8]. The trajectory of the ROI is then tracked and smoothed with a linear Kalman filter.

PRV information in contained in the skin pixels, since its generation is linked to the light reflection by the skin surface. Therefore, it is important to select the skin pixels from the ROI. This is realised by applying the algorithm proposed by Conaire *et al.* in [23].

The RGB color channel values in the area of skin pixels is spatially averaged. This gives a unique RGB triplet per frame. The triplets obtained for the whole video are concatenated to form the RGB temporal traces.

From the existing pulse wave extraction techniques [20][24], the chrominance-based algorithm introduced by De Haan *et al.* in [27] is retained. This algorithm presents analytic formulas, which offers computational simplicity. The RGB traces are first normalized (let us denote R_n, G_n and B_n the normalized traces). Second, two orthogonal signals X_s and Y_s are obtained as:

$$\begin{cases} X_s = 3R_n - 2G_n \\ Y_s = 1.5R_n + G_n - 1.5B_n. \end{cases} \quad (1)$$

A filtering step permits to select pulse frequencies. We use for this a Butterworth filter, with cut-off frequencies of 0.7 and 3.5 HzThe resulting signals X_f and Y_f allow to directly obtain the pulse signal S following equation 2:

$$S = X_f - \alpha Y_f \quad (2)$$

where α is the ratio of X_f and Y_f standard deviations ($\alpha = \frac{\sigma(X_f)}{\sigma(Y_f)}$). This ratio reduces disturbances caused by motion, as they evenly impact X_f and Y_f, contrary to the pulse signal.

The RPPG signal S is interpolated and resampled at 125 Hz, so that the time domain resolution is improved. This makes the pulse wave peak detection easier. The PRV signal is then obtained by calculating the differences of successive pulses occurring times, the PPI, as illustrated in Fig. 1.

The HF and LF components are obtained after PRV waveform power spectrum analysis. The LF component is represented by the spectral power of frequencies between 0.04 Hz and 0.15 Hz, while the HF range covers frequencies between 0.15 Hz and 0.4 Hz. Primary studies of the HRV features proposed the ratio LF/HF as an indicator of the autonomic activity [15]. However, recent researches have proven ambiguities in the interpretation of this ratio [7]. As a consequence, we use the two-dimensional feature (HF, LF) as proposed in [7]. The difference is that we use this representation for the first time for PRV signal analysis whereas it was applied to HRV in [7].

III. Experimental results

A. Dataset presentation

The *Chinese Academy of Sciences Macro-Expressions and Micro-Expressions* (CAS(ME)^2) dataset [28] proposes videos of participants that were filmed while watching emotion-eliciting videos. For each video, the macro and micro facial expressions shown by the subject are listed with their exact occurring times (corresponding frames). It is the first dataset that offers videos with annotated macro and micro-expressions.

Videos that were shown to the subjects induce three different excitation emotions: disgust, anger and happiness. The candidates were asked to control their expressions, in order to facilitate the focus on their appearing micro-expressions, since they are involuntary. 22 participants were filmed, with ages ranging between 19 and 26. The total number of CAS(ME)^2 videos is 97, with lengths of 1 minute to 2min30s. CAS(ME)^2 is originally intended for macro and micro expression spotting and recognition. The current

study is the first to use this dataset for physiological feature extraction.

The macro and micro-expressions are labeled following two ways: the action units analysis (based on the Facial Action Coding System (FACS) introduced by Ekman), and the self-reported emotions from the candidates. There is no annotation for the whole video, except the type of the elicited emotion. Therefore, we consider the type of the emotional excitation as our ground truth when classifying the videos. We need to refer to this annotation because the PRV signals are extracted from an entire video, contrary to macro and micro-expressions which describe punctual facial changes.

B. Results

1) Emotion partitioning based on the (HF, LF) features

We used a *Support Vector Machine* (SVM) classifier to classify the disgust, anger and happiness emotions, based of the two-dimensional (HF, LF) features. The kernel of this SVM is a *Radial Basis Function* (RBF). The obtained classification is illustrated through the (HF, LF) scatter diagram given by Fig. 2. The background colors corresponds to the three emotion classes. The classification was applied to the entire set of the (HF, LF) couples.

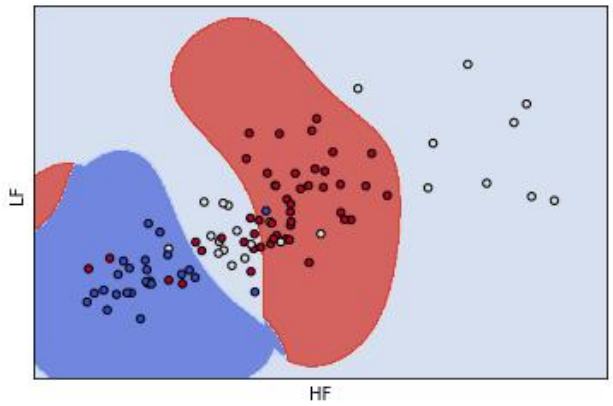


Fig. 2. Resulting emotion partitioning according to the (HF, LF) scatter diagram. Background color correspondence to emotions is: blue for Disgust, red for Anger and clear blue for Happiness

2) Model validation

To validate the configured emotion classifier, the *Leave One Subject Out* (LOSO) is used. In the LOSO protocol, (HF,LF) couples derived from videos of a subject are used as a test set. (HF,LF) values that correspond to the rest of the videos represent the training set. This is an interesting validation method since real-life use confronts the classifier to new subjects.

We obtained the confusion matrix given by Table I:

true \ predicted	*Disgust*	*Anger*	*Happiness*
Disgust	92.0	4.0	4.0
Anger	4.0	79.0	17.0
Happiness	10.0	20.0	69.0

Table I. Emotional classification confusion matrix with the LOSO validation method. Results are expressed in percentage (%)

The accuracy rate of the classification using the LOSO method is 77.32%.

3) Discussion

Emotion classification based on the videos given by CAS(ME)^2 was described qualitatively via visual inspection (Fig. 2) and quantitatively through the LOSO confusion matrix and accuracy rate. The (HF, LF) scatter diagram reveals that disgust and anger form rather distinct sets and can be more easily separated than happiness. This is confirmed by the LOSO validation method, where the disgust matching rate reaches 92.0%. Anger comes in second with 79.0%. Fig. 2 shows that happiness is the most expanded class, leading us to presume a lower matching rate compared to disgust and anger. The confusion matrix illustrated by Table I demonstrates that happiness has effectively the lowest matching score (69.0%).

We notice that disgust is the best classified emotion, which leads us to suggest that the impact of feeling disgust on the sympathovagal activity, and more precisely on the PRV, is greater than anger and happiness. Therefore, the PRV can be a reliable indicator of our reaction to disgust.

IV. Conclusions

This study proposes for the first time physiological signal-based emotion classification using CAS(ME)^2 dataset. Performances using the LOSO validation methods gave interesting results, with an accuracy rate of 77.32%.

This study can be considered as a high-arousal emotion classification; as the disgust, anger and happiness emotions belong to the high arousal domain in the valence-arousal basic emotion classification model. Disgust and anger have negative valence while happiness has positive valence. Further studies can include the 3 other basic emotions, which are fear, sadness and surprise [9].

Since the first aim of CAS(ME)^2 is the recognition and spotting of macro and micro-expression, it is not totally adapted to the PRV extraction. Actually, the participants are not filmed in a neutral emotional state [10]. This would have allowed us to have HF and LF reference values for each subject, in order to normalize the (HF, LF) couples for a better classification and comparison. Working on a dataset that includes a neutral emotional state and more emotional excitation types can be targeted as future works.

V. Ackowledgements

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

VI. References

[1] B. Fasel, J. Lettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 2002.

[2] G. Chanel, J. Kronegg, D. Grandjean et al. Emotion Assessment : Arousal Evaluation Using EEG's and Peripheral Physiological Signals.

[3] G.K. Verma, U.S Tiwary. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage*, 2014.

[4] C. Lisetti, F. Nasoz. Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals. *EURASIP Journal on Applied Signal Processing*, 2004.

[5] S. Jerritta, M. Murugappan, R. Nagarajan et al. Physiological Signals Based Human Emotion Recognition: A Review. *IEEE 7th International Colloquium on Signal Processing*, 2011

[6] J. Wei, H. Luo, S.J. Wu et al. Transdermal Optical Imaging Reveal Basal Stress via Heart Rate Variability Analysis: A Novel Methodology Comparable to Electrocardiography. *Frontiers in Psychology*, 2018.

[7] W. von Rosenberg, T. Chanwimalueang, T. Adjei et al. Resolving Ambiguities in the LF/HF Scatter Plots for the Categorization of Mental and Physical Stress from HRV. *Frontiers in Physiology*, 2017.

[8] P. Viola, M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.

[9] C. Peter, A. Herbon. Emotion representation and physiology assignments in digital systems. *Interacting with Computers*, 2006.

[10] Y. Benezeth, P. Li, R. Macwan et al. Remote heart rate variability for emotional state monitoring. *IEEE International Conference on Biomedical and Health Informatics*, 2018.

[11] E.H. Hon, S.T. Lee. Electronic Evaluation of the Fetal Heart Rate Patterns Preceding Fetal Death, Further Observations. *American Journal of Obstetrics and Gynecology*, 1965.

[12] M. Malik, T. Farrell, T. Cripps et al. Heart rate variability in relation to prognosis after myocardial infarction: selection of optimal processing techniques. *European Heart Journal*, 1989.

[13] D.K. Van Hoogenhuyze, N. Weinstein, G.J. Martin et al. Reproducibility and relation to mean heart rate of heart rate variability in normal subjects and in patients with congestive heart failure secondary to coronary artery disease. *American Journal of Cardiology*, 1991.

[14] R. McCraty, M. Atkinson, D. Tomasino. Science Of The Heart - Exploring the Role of the Heart in Human Performance. *Institute of HeartMath*, 2001

[15] M. Malik, J.T. Bigger, A.J. Camm et al. Heart rate variability - Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 1996.

[16] A. Pichon, D. Chapelot. Horizons in Neuroscience Research Volume 1 - Homeostatic Role of the Parasympathetic Nervous System in Human Behavior. *Nova Science Publishers, Inc.*, 2009.

[17] R. McCraty, M. Atkinson et al. The effects of emotions on short term heart rate variability using power spectrum analysis. *American Journal of Cardiology*, 1995.

[18] J. Liu, H. Luo, P.P. Zheng et al. Transdermal optical imaging revealed different spatiotemporal patterns of facial cardiovascular activities. *Scientific Reports*, 2018.

[19] Y. Sun, S. Hu, V. Azorin-Peris et al. Noncontact imaging photoplethysmography tp effectively access pulse rate variability. *Journal of Biomedical Optics*, 2012.

[20] Y. Sun, N. Thakor. Photoplethysmography revisited: from contact to non contact, from point to imaging. *IEEE Transactions on Biomedical Engineering*, 2016.

[21] E. Gil, M. Orini, R. Bailon et al. Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions. *Physiological Measurement*, 2010.

[22] M. Nitzan, A. Babchenko, B. Khanokh et al. The variability of the photoplethysmographic signal - a potential method for the evaluation of the autonomic nervous system. *Physiological Measurement*, 1998.

[23] C.O. Conaire, N.E. O'Connor, A.F. Smeaton. Detector adaptation by maximising agreement between independant data sources. *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[24] M.Z. Poh, D.J. McDuff, R.W. Picard. Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam. *IEEE Transactions on Biomedical Engineering*, 2011.

[25] D.L. Eckberg, H.H. McGuire. Sympathovagal Balance - A Critical Appraisal. *American Heart Association*, 1997.

[26] G.E. Billman. The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance. *Frontiers in Physiology*, 2013.

[27] G. de Haan, V. Jeanne. Robust pulse-rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering*, 2013.

[28] F. Qu, S.J. Wang, W.J. Yan et al. CAS(ME)^2: A Database for Spontaneous Macro-Expression and Micro-Expression Spotting and Recognition. *IEEE Transactions on Affective Computing*, 2017.

Detection of *H. pylori* induced gastric inflammation by diffuse reflectance analysis

Alexandre Krebs[1], Vania Camilo[2,3], Eliette Touati[2], Yannick Benezeth[1], Franck Marzani[1], Valérie Michel[2], Dominique Lamarque[4], Fan Yang[1]
[1] Univ. de Bourgogne Franche-Comté, LE2I EA 7508, 21078 Dijon, France
[2] Institut Pasteur, Department of Microbiology, Helicobacter Pathogenesis Unit, Paris Cedex 15, France
[3] INSERM U1173, Faculty of Health Sciences, Simone Veil, Université Versailles-Saint Quentin, Saint Quentin en Yvelines, France
[4] Hôpital Ambroise Paré, 92104 Boulogne Billancourt, France
alexandre.krebs@u-bourgogne.fr

Abstract

Spectral acquisitions contain rich information and thus, are promising modalities for early detection of gastric diseases. In this study, we analyze the diffuse reflectance of the gastric inflammatory lesions induced by the bacterium H. pylori in the mouse stomach. We have designed a pipeline to characterize and classify spectra acquired on mice. The pipeline is based on a band clustering algorithm and the computation of meaningful division and subtraction features for classification. Currently, the pipeline is able to recognize inflamed stomachs spectra with an accuracy of 98%. These results are promising and the same pipeline could be adapted for the study of gastric pathologies in humans.

1. Introduction

Spectral acquisitions have been proposed as a useful tool to provide information on the lesions of a tissue and could be promising for the early detection of human diseases. Ogihara *et al.* and Goto *et al.* have used spectral information to diagnose cancer lesions in humans. After intra-patient normalization, it appeared that spectra of gastric tumors differ in the spectral range 650 nm - 800 nm. Moreover, the correlation between spectral reflectance and micro-vascular density is higher in this wavelength range [1], [2]. Akbari *et al.* have also shown their interest on gastric cancer in humans. They have classified spectra using a normalized cancer index, calculated according to the reflectance value in the infrared spectral range (1200 nm - 1400 nm) [3].

Other optical devices have been used to characterize normal gastric tissue and diseased gastric tissue. Grosberg *et al.* use hyperspectral two-photon microscopy for characterization and unmixing of gastric tissues [4]. Bergholt *et al.* have acquired spectra thanks to Raman spectroscopy. These spectra are unmixed to find the relative spectral contribution of DNA, protein, lipid, blood and glycoprotein. The contributions were used to recognize four states: normal gastric tissue, intestinal metaplasia, dysplasia and adenocarcinoma [5]. Cancer can also be detected thanks to fluorescence: Martin *et al.* have developed an hyperspectral imaging system to detect cancer on mice. They have used the wavelengths between 420 nm and 480 nm as excitation to acquire spectra between 530 nm and 580 nm. Ratios and subtractions between spectra were used to distinguish malignant tissue from normal one [6].

Gastric cancer is the third most common cause of cancer-related deaths in the world according to GLOBOCAN 2012 [7]. Gastric cancer is often associated with a bad prognosis mainly due to its detection at an advanced stage. The current practice to detect gastric lesions on humans is to conduct biopsies, *i.e.* to remove a piece of stomach and examine it in laboratory. The analysis of reflectance has the advantage to be non-invasive for the tissue compared to biopsies and thus, it is a convenient mean to characterize gastric lesions without damaging the tissues. The infection by *Helicobacter pylori* is the major cause of development of gastric cancer that has been identified. This bacterium which colonizes specifically the human stomach is responsible for the most common infection among the worldwide population. *H. pylori* is considered as a class I carcinogen by the World Health Organization [8]. More precisely, *H. pylori* is able to adapt to the gastric acidity and to persistently colonize the gastric mucosa, affect precursor and stem cells [9]. *H. pylori* is associated with various gastric diseases in humans. It induces a chronic gastritis

that can evolved to ulcer diseases or led to intestinal metaplasia, dysplasia and gastric cancer [10] [11]. It has also been shown that *H. pylori* induces a gastric mutagenic effect in mice, associated with the induced-inflammation of the stomach**touati2003**

In order to investigate the ability of spectral reflectance to distinguish between normal and inflamed tissue, in this study we compare the spectral reflectance of the gastric mucosa of chronically *H. pylori* -infected mice to non-infected controls. The data led to design a defined pipeline to classify the obtained spectra between two classes: "Normal" and "Inflamed mucosa".

The rest of the paper is organized as follows. The protocol used to infect mice and to acquire spectra is introduced in section II. The method used to process and classify spectra is detailed in section III. Classification results are discussed in section IV.

2. Protocol and acquisitions

Mice experiments were carried out in strict accordance with the recommendations in the Specific Guide for the Care and Use of Laboratory Animals of the Institut Pasteur, according to the European Directives (2010/63/UE). The project was approved by the Committee of Central Animal Facility Board of the Institut Pasteur and by the Comité d´Ethique en Expérimentation Animale (CETEA), Institut Pasteur (Ref 2013-0051). Five-six weeks old specific pathogens free C57BL/6 male mice (Charle River, France) (n=72) are used. Half of mice (n=36) were orogastrically infected with 100 μl of the *H. pylori* mouse-adapted strain SS1 (10^8 colonies forming unit (cfu)/ml). The SS1 strain is able to colonize chronically the mice stomach for several months, leading to an active gastritis [12]. The local and systemic immune response on *H. pylori* SS1-infected mice are similar to those described in human infections [13]. One control group, serving as reference and one group orally infected by *H. pylori* strain SS1. [14]. The non-infected group of mice (n=36) received orogastrically 100 μl of peptone broth as previously described **touati2003** After 1, 3, 6, 12 and 18 months, mice are sacrificed and stomach isolated for reflectance spectral acquisition, quantification of *H. pylori* gastric colonization and histological analysis of the gastric lesions. In addiction, blood was collected for the determination of *H. pylori* serology.

En cours.... at several time-points after infection: 1, 3, 6, 12 and 18 months (Acquisitions at 1,3 and 6 months are out of scope of the present study). In total 34 mice were sacrificed (18 non-infected and 16 infected). Among non-infected mice, four of them have still developed an inflammation while two infected mice do not show any sign of inflammation while being infected. The distribution of non-infected and infected mice according to their histology report are presented on Figure 1.

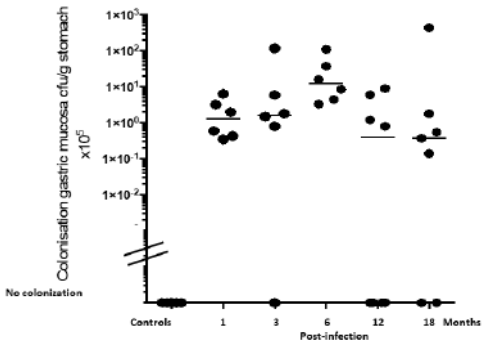


Figure 1. Number of mice per class

After sacrifice, the stomach is then placed between coverslips and spectra are acquired on the glandular side by a spectrometer. The spectrometer comes from the Avantes company and is able to acquire spectra between 200 nm and 1160 nm with non-constant spectral resolution (reference: AvaSpec-ULS2048XL-EVO). To provide a diffuse illumination on the tissue sample, an integrating sphere, combined with an halogen light source is used (reference: AvaSphere-50-LS-HAL-12V from Avantes). About five spectra per mice are acquired on different locations. The spectra are obtained by averaging over the sample port of 1cm diameter.

3. Method

The obtained spectra could be noisy for some wavelengths, thus, only measures between 400 nm and 950 nm are kept. Still, this spectral range corresponds to 953 bands of the spectrometer. This number is too large compared to the number of spectra. The use of all bands would lead up to overfitting problems in the classification process. Thus, an algorithm has been designed to merge neighboring bands and create a reduced set of bands. The strategy is to compute recursively the correlation between neighboring bands and merge the most correlated couple of bands. The algorithm stops when the highest correlation found is under a certain threshold. In practice we use a threshold $\tau = 0.95$, this corresponds to a reduction from 953 bands to 6 groups of bands as presented on Figure 2. The reduction seems rough but this process keeps most of the information.

Figure 3 shows the six resulting wavelengths groups and the mean spectra of mice. Let λ_i be a cutoff wavelength with λ_{0-6} = 400 nm, 451 nm, 516 nm, 589 nm, 620 nm, 668 nm and 950 nm. The last group between 668 nm and 950 nm is relatively large. This indicates that all wavelengths in this range are highly correlated between them and information contained in this range are redundant.

The bands that have been grouped are reduced to their mean μ_i over the considered wavelength range as in equa-

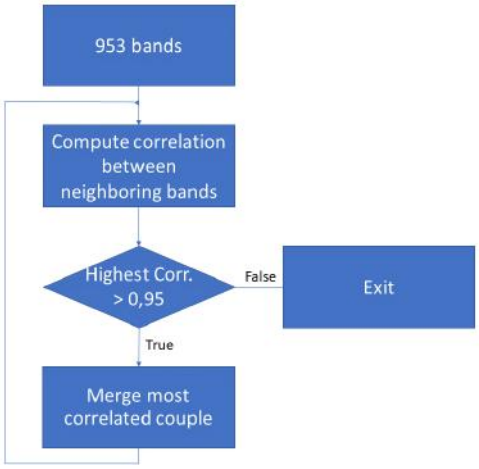


Figure 2. Band Reduction Algorithm

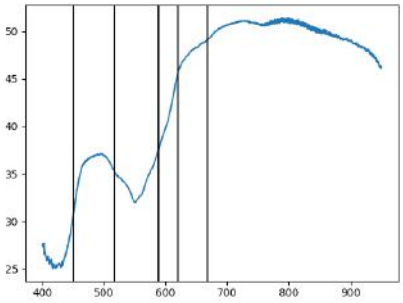


Figure 3. Separation between the six groups of wavelengths

tion 1:

$$\mu_i = \frac{1}{N_i} \sum_{\lambda=\lambda_i}^{\lambda_{i+1}} S_\lambda \quad (1)$$

where N_i is the number of wavelengths contained in the range $[\lambda_i,\lambda_{i+1}[$ and S_λ is the reflectance of a spectra at wavelength λ.

This algorithm shares similarities with agglomerative clustering. At the beginning of the algorithm, agglomerative clustering considers each element as one cluster and merge them recursively. A version of this algorithm works with connectivity constraints [15]. In our case this connectivity constraint is that only two neighboring bands can be merged. As for agglomerative clustering, this algorithm is unsupervised in the sense that the label of each mice is not used to find optimal cuts. This choice is essential to avoid introducing bias in the following classification process.

After reducing the number of wavelengths, features are extracted. In practice, we can use either directly the reduced bands as features either engineering other types of features that are meaningful for classification. In the present case, "division" features and "subtraction" features have been investigated *i.e.* every possible subtraction or division between the six group of bands are computed. The key idea is that these features are respectively insensitive to offset and scaling of the spectra and thus are more robust for classification. More formally, we denote the subtraction features and the division features as in equations 2 and 3:

$$D_{i,j} = \lambda_i - \lambda_j; \quad (2)$$

$$R_{i,j} = \frac{\lambda_i}{\lambda_j} \quad (3)$$

A step of univariate feature selection is then performed to reduce the complexity of the classifier and to enhance generalization by reducing overfitting. ANalysis Of VAriance (ANOVA) is used to select the *k* most discriminative features. Classification is then applied on selected features using a classical Support Vector Machine (SVM) with linear kernel. A classifier is trained on a Leave One Out (LOO) per mouse loop. Meaning that, on each iteration, all spectra are used for training the classifier except spectra from one mouse. The classifier is tested on this last mouse. The label of each mouse is defined by its histological class *i.e.* mice with histology (-) or ($\pm$) are labeled "normal" and others (+, ++, +++, ++++) are labeled "inflamed" in connection with histological tables.

The full classification pipeline is designed in Python programming with Scikit-learn library [16] and is summarized on Figure 4.

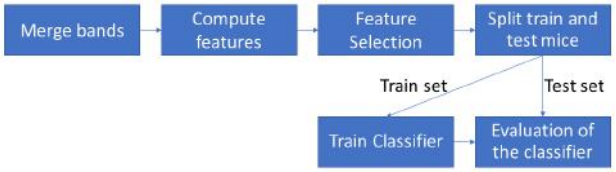


Figure 4. Classification Pipeline

4. Results and discussion

As data at 1, 3 and 6 months were not significant, we have decided to use mice from 12 months and 18 months data to build the classifier and the three types of features have been tested: The reduced bands (*ID*), the division features (*Div*) and the subtraction features (*Sub*). Figures 5, 6, 7 present the statistical distribution of the features according to sample's label. Red boxplots are the inflamed mice and blue boxplots are not-inflamed mice. Figure 5 shows the distribution for *ID* features. It can be seen that μ_0 and μ_5 are not informative for classification because red and blue boxplots are almost identical. On the contrary, μ_1, μ_2, μ_3 and μ_4 are more informative. In the same way, Figure 6 shows that features $R_{1,2}$, $R_{1,3}$ and $R_{2,3}$ are not informative but $R_{3,5}$ and $R_{4,5}$ are very informative. Finally, boxplots

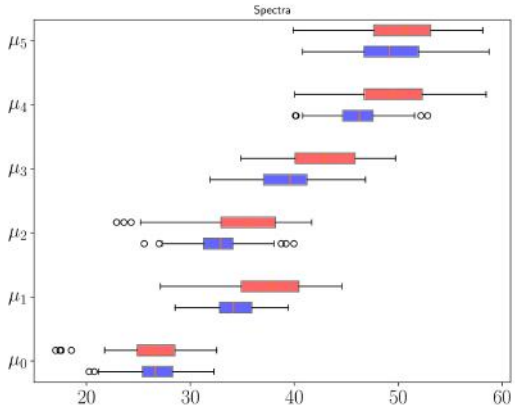


Figure 5. Distribution of *ID* features

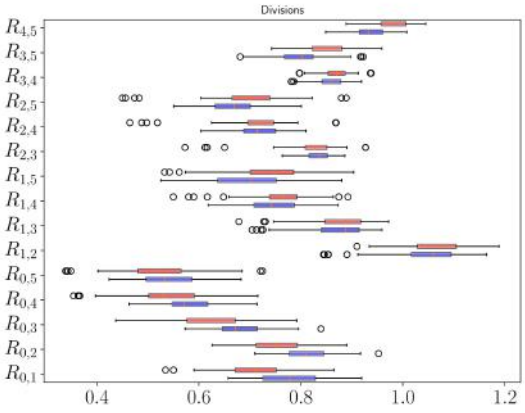


Figure 6. Distribution of *Div* features

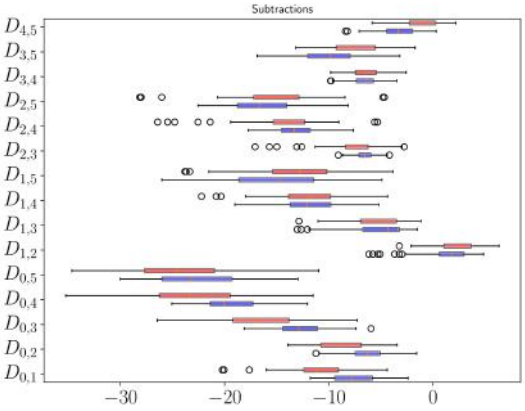


Figure 7. Distribution of *Sub* features

on Figure 7 show that features $D_{3,4}$, $D_{1,4}$, $D_{1,3}$ are ineffective for classification while $D_{0,1}$ and $D_{4,5}$ are interesting to discriminate between the two classes.

More concrete, Table 1 summarizes the classification results obtained with the SVM classifier. The best results obtained with ID features were 79% precision, 76% recall and a F1 score of 76%. This result is close to the result given by Sub. features. This is expected because linear SVM try to find the best hyperplane that separate data. Thus a linear transformation on data (*i.e.* subtractions) will induce the same hyperplane. On the opposite, division features give better results: 98% accuracy, 98% recall and 98% as F1 score. This shows the relevancy of computing ratios between wavelengths (or in our case, between groups of wavelengths). Ratios have the advantage to be scaling invariant and thus no normalization is needed.

Table 1. Classification Results

	Precision	Recall	F1 score
ID	0.79	0.76	0.76
Div	0.98	0.98	0.98
Sub	0.81	0.77	0.77

Moreover, the results presented on table 1 are the best results among all tests by varying the number of features selected. Divisions result, with 98% of precision is obtained thanks to a single ratio: Most of the time the ratio $R_{4,5}$ (*i.e.* the ratio between groups [620 nm, 668 nm[and [668 nm, 950 nm[) was chosen to be the most discriminative feature. These two wavelength ranges seem very interesting for inflammation detection and basing the classification on this single ratio make it simpler, more robust and with a high power of generalization.

5. Conclusion

In this paper, a method of characterization and classification of stomach spectra is introduced. The classification pipeline is able to recognize an inflamed stomach reflectance spectrum from a normal one with an accuracy of 98%. Moreover we have proven that some ratios between groups of bands were interesting features for the classification. Spectral range [620 nm, 668 nm[and [668 nm, 950 nm[are the most useful to discriminate normal gastric tissue from inflamed tissue. As a perspective, the same pipeline could be reused and adapted to gastric cancer detection on human and a simplified device, focused on specific wavelength range can be built to help practicians in their daily work. Moreover, adding spatial information by using multispectral imaging could combine characterization and location of gastric diseases.

Acknowledgment

This study was supported by the French Research National Agency (ANR) program EMMIE under the grant agreement 15-CE17-0015.

References

[1] A. Goto, J. Nishikawa, S. Kiyotoki, M. Nakamura, J. Nishimura, T. Okamoto, H. Ogihara, Y. Fujita, Y. Hamamoto, and I. Sakaida, "Use of hyperspectral imaging technology to develop a diagnostic support system for gastric cancer," *Journal of biomedical optics*, vol. 20, no. 1, p. 016 017, 2015 (cit. on p. 1).

[2] H. Ogihara, Y. Hamamoto, Y. Fujita, A. Goto, J. Nishikawa, and I. Sakaida, "Development of a gastric cancer diagnostic support system with a pattern recognition method using a hyperspectral camera," *Journal of Sensors*, vol. 2016, 2016 (cit. on p. 1).

[3] H. Akbari, K. Uto, Y. Kosugi, K. Kojima, and N. Tanaka, "Cancer detection using infrared hyperspectral imaging," *Cancer science*, vol. 102, no. 4, pp. 852–857, 2011 (cit. on p. 1).

[4] L. E. Grosberg, A. J. Radosevich, S. Asfaha, T. C. Wang, and E. M. Hillman, "Spectral characterization and unmixing of intrinsic contrast in intact normal and diseased gastric tissues using hyperspectral two-photon microscopy," *PLoS one*, vol. 6, no. 5, e19925, 2011 (cit. on p. 1).

[5] M. S. Bergholt, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. Y. So, A. Shabbir, and Z. Huang, "Fiber-optic raman spectroscopy probes gastric carcinogenesis in vivo at endoscopy," *Journal of biophotonics*, vol. 6, no. 1, pp. 49–59, 2013 (cit. on p. 1).

[6] M. E. Martin, M. B. Wabuyele, K. Chen, P. Kasili, M. Panjehpour, M. Phan, B. Overholt, G. Cunningham, D. Wilson, R. C. DeNovo, *et al.*, "Development of an advanced hyperspectral imaging (hsi) system with applications for cancer detection," *Annals of biomedical engineering*, vol. 34, no. 6, pp. 1061–1068, 2006 (cit. on p. 1).

[7] J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J.-W. W. Coebergh, H. Comber, D. Forman, and F. Bray, "Cancer incidence and mortality patterns in europe: Estimates for 40 countries in 2012," *European journal of cancer*, vol. 49, no. 6, pp. 1374–1403, 2013 (cit. on p. 1).

[8] J. M. Noto and R. M. Peek, "Helicobacter pylori: An overview," in *Helicobacter Species*, Springer, 2012, pp. 7–10 (cit. on p. 1).

[9] M. Amieva and R. M. Peek Jr, "Pathobiology of helicobacter pylori–induced gastric cancer," *Gastroenterology*, vol. 150, no. 1, pp. 64–78, 2016 (cit. on p. 1).

[10] P. Correa, "Human gastric carcinogenesis: A multistep and multifactorial processfffdfffdfffdfirst american cancer society award lecture on cancer epidemiology and prevention," *Cancer research*, vol. 52, no. 24, pp. 6735–6740, 1992 (cit. on p. 2).

[11] M. D. Burkitt, C. A. Duckworth, J. M. Williams, and D. M. Pritchard, "Helicobacter pylori-induced gastric pathology: Insights from in vivo and ex vivo models," *Disease models & mechanisms*, vol. 10, no. 2, pp. 89–104, 2017 (cit. on p. 2).

[12] A. Lee, J. O'Rourke, M. C. De Ungria, B. Robertson, G. Daskalopoulos, and M. F. Dixon, "A standardized mouse model of helicobacter pylori infection: Introducing the sydney strain," *Gastroenterology*, vol. 112, no. 4, pp. 1386–1397, 1997 (cit. on p. 2).

[13] R. L. Ferrero, J.-M. Thiberge, M. Huerre, and A. Labigne, "Immune responses of specific-pathogen-free mice to chronic helicobacter pylori (strain ss1) infection," *Infection and immunity*, vol. 66, no. 4, pp. 1349–1355, 1998 (cit. on p. 2).

[14] A. S. Day, N. L. Jones, Z. Policova, H. A. Jennings, E. K. Yau, P. Shannon, A. W. Neumann, and P. M. Sherman, "Characterization of virulence factors of mouse-adapted helicobacter pylori strain ss1 and effects on gastric hydrophobicity," *Digestive diseases and sciences*, vol. 46, no. 9, pp. 1943–1951, 2001 (cit. on p. 2).

[15] J. Li, "Agglomerative connectivity constrained clustering for image segmentation," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 4, no. 1, pp. 84–99, 2011 (cit. on p. 3).

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011 (cit. on p. 3).

Outdoor Scenes Pixel-Wise Semantic Segmentation using Polarimetry and Fully Convolutional Network

Marc Blanchon[1], Olivier Morel[1], Yifei Zhang[1], Ralph Seulin[1], Nathan Crombez[2] and Désiré Sidibé[1]
[1]*ImViA EA 7535, ERL VIBOT CNRS 6000, Université de Bourgogne Franche Comté (UBFC), 12 Rue de la Fonderie, 71200, Le Creusot, France*
[2]*EPAN Research Group, University of Technology of Belfort-Montbéliard (UTBM), 90010, Belfort, France*
marc_blanchon@etu.u-bourgogne.fr

Keywords: polarimetry, deep learning, segmentation, augmentation, reflective areas.

Abstract: In this paper, we propose a novel method for pixel-wise scene segmentation application using polarimetry. To address the difficulty of detecting highly reflective areas such as water and windows, we use the angle and degree of polarization of these areas, obtained by processing images from a polarimetric camera. A deep learning framework, based on encoder-decoder architecture, is used for the segmentation of regions of interest. Different methods of augmentation have been developed to obtain a sufficient amount of data, while preserving the physical properties of the polarimetric images. Moreover, we introduce a new dataset comprising both RGB and polarimetric images with manual ground truth annotations for seven different classes. Experimental results on this dataset, show that deep learning can benefit from polarimetry and obtain better segmentation results compared to RGB modality. In particular, we obtain an improvement of 38.35% and 22.92% in the accuracy for segmenting windows and cars respectively.

1 INTRODUCTION

Scene segmentation and understanding have been a popular topic in the field of robotics, artificial intelligence and computer vision. It has attracted a lot of research with different aproaches: decision forest approach (Gupta et al., 2014), deep approach for semantic segmentation (Couprie et al., 2013), and pixel-wise semantic segmentation (Badrinarayanan et al., 2015). The main challenge lies in the recognition and the assignment of multiple classes.
A difficult key point when addressing the problem of segmentation is the possible presence of reflective areas. The segmentation method should be able to differentiate a physical object and its projection on a reflective area.

The field of segmentation of complex scenes is open since many applications could benefit. Some research has been conducted on the detection of mud (Rankin and Matthies, 2010a), as well as on the detection of water (Yan, 2014; Nguyen et al., 2017). Indeed, robotics and autonomous cars could take advantage of these abilities. For example, if a system is able to understand a scene with complex areas (reflective), then it is possible to avoid them.

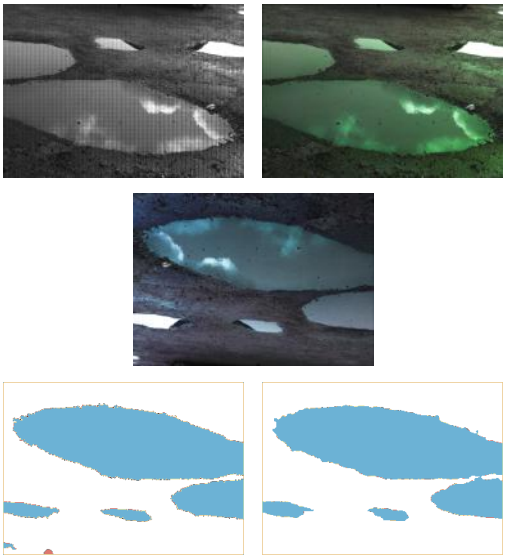


Figure 1: From raw polarimetric image to segmentation. Top: left is the raw polarimetric image, right is the transformed image to HSL (Hue Saturation Luminance). Middle image is the augmented image with proper physical meaning. Bottom: left is the hand made ground truth and right is the prediction of the deep learning network for the middle image.

To handle both the classification of so-called standard zones (or "low complexity") and areas of high complexity, the introduction of a discriminant modality is considered.
The choice is oriented towards the polarimetric imaging, giving the ability to measure and recover the changes in the light waves. SFP (shape-from-polarization) techniques have been using the ability of polarimetry to extract information from highly reflective objects (Rahmann and Canterakis, 2001; Morel et al., 2005). Therefore, polarimetric cameras have experienced a big development leading to better ease of use and practicality. The Division of Focal Plane (DoFP) allows the capture of an image using four different polarizers. In consequence, it is similar as acquiring four images with four polarizers.
Combining the advantages of different data types, a polarimetric camera will process non-reflective data as usual gray-scale portion of the image, while reflective areas will observe changes in the image information. In consequence of using polarimetric images, a set of constraints has been deduced to design a data augmentation process.
Since the aim of this paper is to measure and qualify the usefulness of a complex modality applied to a specific task, it is unnecessary to complexify the task at the early stage of the processing. Consequently, a widely used and tested network is the core of this study: SegNet (Badrinarayanan et al., 2015). The robustness and modularity of this architecture makes this network the perfect candidate for our purpose.
As shown in Figure 1, this paper allows understanding and exploitation of this new type of information in the context of deep learning.
This paper proposes the following main contributions:

- Introduction of the polarimetry in the field of feature learning to discuss the advantages and disadvantages of such data. In addition, a dataset has been created for the experimental needs.
- Creation of novel techniques allowing polarimetric data to be augmented by preserving the physical properties from this modality.
- Detection and segmentation of reflective areas through standard convolutional deep learning techniques.

The various past works on which this paper is based are presented in Section 2. Then, the different processes of our implementation are introduced in the Section 3. The forth section summarize all the necessary steps for the experiment. Also, this section presents the results of the two modalities used (polarimetry and RGB) and the discussions that will compare the results obtained and also their interpretations. The last section concludes on this work as well as offers an opening on future work.

2 RELATED WORKS

2.1 Scene Segmentation

The pixel-wise semantic segmentation is the ability of giving a label for each pixel of an image. This task requires an accurate learning of the features on a set of image. This leads to the creation of a generic model which is able to classify at the pixel-level. Many research proved that deep learning models tend to make complex task learning and understanding accessible. Computer vision has benefited from the advances in this field to progress in general tasks. More precisely, many applications of semantic segmentation has been developed; among the most represented: road scene segmentation (Oliveira et al., 2016), indoor scene understanding (Gupta et al., 2014; Qi et al., 2017).

The first remarkable deep learning based segmentation is the FCN from Long et al. (Long et al., 2015), that allows the segmentation of image of any sizes without fully connected layers. Starting from this previous paper, as the years and the evolution of power increased, multiple networks, each with better performance, have been released: SegNet (Badrinarayanan et al., 2015), DeepLab (Chen et al., 2015; Chen et al., 2016; Chen et al., 2018), Image-to-Image (Isola et al., 2017) , Conditional Generative Adversarial Networks (Wang et al., 2018).

2.2 Polarimetry

Polarimetry is the science of measuring the polarized state of the light. As a consequence, a polarimetric camera (Wolff and Andreou, 1995) gives the experience of recovering the light changes in the captured environment. Because of this behavior, the information from this camera could be the perfect candidate as a discriminant factor for complex scene semantic understanding.
As shown in Figure 2, polarimetric images can be used advantageously, because the reflection operates a direct impact on the image.
For example, Kai Berger et al. proposed a method for depth recovering from polarimetric images in urban environment (Berger et al., 2017), treating the modality as a common RGB camera. Other polarization based systems have been proposed for water detection using polarized information. For example, Nguyen et

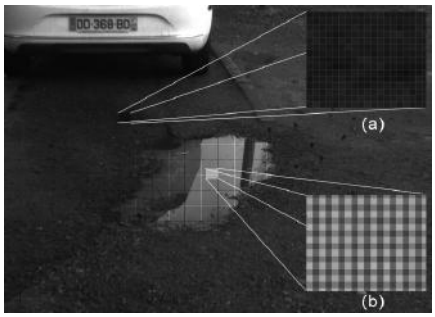


Figure 2: Reflection Influence on Polarimetry. (a) is a zoom on the non-polarized area and (b) on a polarized area. Clearly, on a polarized surface, the micro-grid appears and reveals an intensity change according to the polarizer affected.

al. proposed a method for water tracking with a polarized stereo system (Nguyen et al., 2017) achieving an approximate accuracy of 65% exceeding the previous state of the art method accuracy of approximately 45% (Yan, 2014). Rankin and Matthies proposed an application in recognition of mud for autonomous robotics and offered a full benchmark for the segmentation processes (Rankin and Matthies, 2010b).
One of the disadvantages of these previous methods is the lack of automation of tasks or the difficulty of deployment. In contrast, a deep learning approach allows the creation of a model that can be reused and redesigned as it goes along.

Despite the useful and informative aspects of polarimetric system, the use of such cameras have been quite restricted, due to the limitation of hardware and automatic integration. Using the DoFP technique (Nordin et al., 1999b; Nordin et al., 1999a; Millerd et al., 2006), the polarimetric camera has been introduced, which allow easier integration. DoFP technique allows having the polarized filters in an array directly on the sensor. In this design, four polarized filters, with unique angles, are used to capture four different measurements instantly in one shot. Many image processing and computer vision applications can benefit from recent DoFP-polarimetric camera.

In this paper, we are introducing polarimetry to the field of pixel-wise semantic segmentation for outdoor scenes.

3 METHOD

3.1 Polarimetric data pre-processing

Contrary to other standard type of images (RGB, gray-scale, etc.), the image provided by a DoFP camera is composed of 2x2 super-pixels. Consequently, we use an interpolation method (Ratliff et al., 2009) in order to recover polarimetry images. The key idea behind this transformation is to extract three one-channel images to represent three physical notions: the Angle of Polarization (AoP), the Degree of Polarization (DoP) and the Intensity (I). The AoP represents the value of the angle of polarization at each pixel while the DoP is the strength of the polarization state of the incoming light for each pixel.
In nature, the light is mainly partially linearly polarized which reduces the Stokes parameters to three parameters as bellow:

$$S = \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ 0 \end{pmatrix} = \begin{pmatrix} P_0 + P_{90} \\ P_0 - P_{90} \\ P_{45} - P_{135} \\ 0 \end{pmatrix}, \tag{1}$$

where $s_{\{0,1,2\}}$ are the three-first Stokes parameters, and $P_{\{0,45,90,135\}}$ the intensity output images corresponding to the orientation of the polarizer. The commonly used Stokes vectors can be normalized by s_0:

$$\bar{S} = \begin{pmatrix} \bar{s_1} \\ \bar{s_2} \\ 0 \end{pmatrix} = \frac{1}{s_0} \begin{pmatrix} s_1 \\ s_2 \\ 0 \end{pmatrix}. \tag{2}$$

AoP and DoP can be deduced according to:

$$\text{DoP} = \sqrt{\bar{s_1}^2 + \bar{s_2}^2}, \tag{3}$$

$$\text{AoP} = \frac{1}{2}\tan^{-1}\left(\frac{s_1}{s_2}\right). \tag{4}$$

The last parameter I is the intensity which is the combination of all polarized states intensities:

$$\text{I} = \frac{P_0 + P_{45} + P_{90} + P_{135}}{2}. \tag{5}$$

After this computation, three gray-scale description images of the raw polarimetric data are obtained. We have chosen to build an HSL (Hue Saturation Luminance) image mapping the three previous sources of information. This colorspace allows specific behavior per channel which fit with the data provided by AoP, DoP and I. The hue is commonly a 360° periodic value, the saturation is a value between zero and one as well as the value for the luminance. To fit the prerequisites of this color space, we made the adaptation and/or normalization of our images according to each channel and then merged them together (Wolff and Andreou, 1995).

$$H \longrightarrow 2 * \text{AoP}, \quad S \longrightarrow \text{DoP}, \quad L \longrightarrow I/255. \tag{6}$$

HSL can be seen as a single 3-channel image. This allows any RGB pre-initialized DL network to deal with these images. It is then possible to augment the data taking advantage of the HSL representation.

3.2 Polarimetric data augmentation

As previously explained, polarimetric information characterizes the vectorial representation of light. By consequence, any image has a unique meaning only for these precise camera parameters and orientations. The augmentation procedure consists in creating new images with the application of a transformation and/or an interpolation. The constraints induced by the type of data are exported to any transformation applied. The luminance and saturation channels can be released of the constraints because their attributed values are invariant around the optical axis. Contrarily, the hue is affected by this transformation. It is necessary to recompute the hue coherently with the physical properties of the camera. In this unique case, the angle of polarization will have a consistent physical meaning.

While rotating the camera counter-clockwise, the angle of polarization is rotated clockwise. Let θ be the applied rotation angle to the camera, R_θ the rotation matrix and H the hue channel of the image:

$$\mathrm{H}_{\text{rotated}} = R_\theta(\mathrm{H}_{prev} - \theta). \tag{7}$$

At the end of this computation, the image will keep its physical properties and be rotated.
As shown in Table 1, a set of transformations has been developed to give the ability to extend any polarimetric images dataset and it is remarkable that only the hue channel needs some modifications to stick to physical properties. The translation is only a shift in the images, which means that there is no modification in the view point of the camera. Since a polarimetric camera is dependent on the actual position and view point, the hue channel remains invariant to translation. On the other hand, if the camera lens has a wide angle, then in this case an additional transformation will be necessary (Table 1 -*).

3.3 Pixel-wise Segmentation with Deep Learning

Deep learning shows great performances on learning new kind of features and giving genericity to a model.

SegNet (Badrinarayanan et al., 2015) is employed in our work because of its robustness and short training time. The SegNet has an encoder-decoder design and an architecture composed of 36 layers. In our application, the key point in this design lies in the encoder part. It is composed of 13 layers, fitting perfectly the VGG-16 (Simonyan and Zisserman, 2014) ConvNet configuration B. In consequence, a transfer learning (Pan et al., 2010; Torrey and Shavlik, 2010) method can be applied allowing pre-initialization of the network. Considering this approach, an efficient training can be operated, avoiding a costly end-to-end training.

3.4 A New Dataset: PolaBot

Acquisition was conducted to provide a new multi-modal dataset PolaBot with polarimetric images. To the best of our knowledge, no such specific dataset has been released yet. Moreover, in order to make this dataset reliable for different fields (robotic, autonomous navigation, etc.), the acquisitions were made with a multi-modal system of four calibrated cameras. Three synchronized modalities are represented, two RGB from different angles, one NIR (Near-Infrared) and one polarimetric camera. In addition, this collection of information will allow a strong and efficient benchmark, giving the opportunity to compare standard modality to the polarimetry for the exact same scenes and application. This dataset is available at: http://vibot.cnrs.fr/polabot.html.

4 EXPERIMENTS

To confirm our hypothesis of the polarimetric data being more efficient than standard modality for our application, experiments have been conducted, allowing a comparison.
All the experiments were performed on the same dedicated server composed of an Nvidia Titan Xp (12GB Memory) GPU, 128GB of RAM and two CPU accumulating a total of 24 physical cores (48 threads).

For the SegNet Network, internal parameters of the training must be set. We had to set the loss function and the optimizer. We decided to use Adam (Kingma and Ba, 2014) as optimizer and as the loss function the cross entropy loss, defined as:

$$CEL(p,q) = -\sum_{\forall x} p(x) log(q(x)), \tag{8}$$

where x represents the class, $p(x)$ is the prediction for the x class and $q(x)$ the ground truth. Also, for all the training, the learning rate was initialized as 10^{-4} and a maximum of 500 epochs.

Table 1: Augmentation procedure per channels. Here "-" represents invariant, "*" represents that under condition this parameter can be modified.

	AoP (H)	DoP (S)	Intensity (L)
Crop	-	-	-
Roation	$R_\theta(\mathrm{H}-\theta) \pmod{360}$	-	-
Symmetry (Flip)	$-H \pmod{360}$	-	-
Translation	-*	-*	-

4.1 Metrics

To measure the efficiency of the training, common metric has been employed during the process: MIoU (Mean Intersection over Union), F1 Score, Mean Accuracy and Overall Accuracy. The IoU is defined as:

$$\mathrm{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}}. \quad (9)$$

Another widely used metric is the F1 score. This metric observes the same behavior as the MIoU since the perfect score is 1. This metric is a combination of the recall and the precision, which correspond respectively to the relevance and the robustness of the results:

$$\text{F1 Score} = 2 * \frac{\text{precision . recall}}{\text{precision} + \text{recall}}. \quad (10)$$

Finally, the per-class accuracy is the measurement of fitting for each class:

$$\text{Accuracy}_C = \frac{\sum_i [\, p(i) = C \cap GT(i) = C \,]}{\sum_i [\, GT(i) = C \,]}, \quad (11)$$

where C is the class, $p(i)$ is the predicted class of pixel i and $GT(i)$ the ground truth.

4.2 Results

A color chart is used, therefore, for the next images, each area color in the image will have a meaning shown in the Table 2.

Each class has a clear meaning except unlabeled and None. None corresponds to zones segmented by hand but considered non-revealing with respect to our application. The unlabeled class, on the other hand, comes from manual segmentation errors. This class is the eighth class but is not necessarily consistent. Therefore, the results for this class will be neglected and taken into account in the conclusions drawn.

4.2.1 Training Results

Metrics for each epoch has been computed. This procedure allows seeing the fitness evolution of the model.

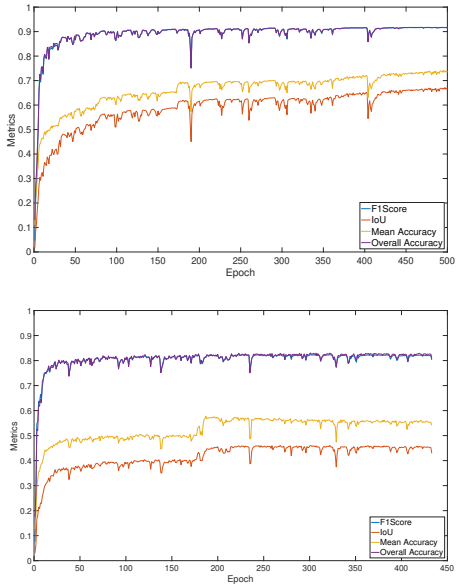


Figure 3: Training Results - Top is the graph corresponding to metrics estimation for the polarimetric data while training. The bottom graph corresponds to the RGB data training.

As shown in Figure 3, both curves are different according to the data provided to the network. First, it is possible to see that the two processes did not stop at the same time. While the network with polarimetric data reached 500 epochs, the network processing RGB data ended at 432 epochs. Indeed, we had put into place a stopping criterion to prevent the network from decaying. This result means that the SegNet RGB has experienced a decrease in its validation metrics for more than 10 epochs. However, our process allows the recovery of the optimal state in order to assess the so-called "optimal" results.
In a second step, it is possible to notice the differences in metric values. The SegNet Polarimetry reaches a MIoU value of 0.66, an F1 score of 0.91 and an average accuracy of 0.73. On the other hand, the SegNet RGB appears to be less efficient with lower scores:

Table 2: Color chart. This color chart allows uniformity in the visualization of results (each class has an affiliated color).

Meaning	Unlabeled	Sky	Water	Windows	Road	Cars	Building	None
Color	Black	Green	Blue	Yellow	Orange	Red	Grey	White

MIoU of 0.42, F1 score of 0.8 and average accuracy of 0.54.
It is possible to conclude this estimate of training by stating that SegNet Polarimetry seems to perform better during the learning phase.

4.2.2 Testing Results

The testing results correspond to the results obtained at the output of the network. As shown in the Table 3, in order to compare the impacts of each type of data, their respective accuracy by class was calculated for RGB and polarimetry and followed by comparison via difference:

$$\text{Accuracy}_{\text{Diff}} = \text{Accuracy}_{\text{Pol}} - \text{Accuracy}_{\text{RGB}}. \quad (12)$$

The Figure 4 shows the results obtained at the output of the SegNet Polarimetry and the Figure 5 those of the SegNet RGB. The segmentation is correct in both cases and visually offers good results.

4.3 Discussion

As shown in the Table 3, very high accuracy can be observed in all segmented classes using polarimetric data. As the data set is not generic, the sky remains on the same tone (blue), which gives a significant advantage over the RGB mode. The other classes where the RGB model is better are: road, water and none. These differences are minimal and can be explained in several ways. One of our hypotheses concerns the difference in manual segmentation for ground truth. RGB and polarimetry were segmented independently, increasing uncertainties. The difficulty of segmentation of certain classes must be taken into account. Another way to look at these results is to consider the advantages and disadvantages of cameras in relation to the dataset. For example, the road can be polarized if there is a high temperature; therefore, polarimetry would have an advantage over the RGB model. Since the dataset is acquired in only one type of weather condition, the RGB may have an advance over the other model, which may explain these results.
However, polarimetry model gives very high accuracy in all the classes. More precisely, when segmenting areas such as windows, cars and building, the model obtain a big positive difference compared to the RGB. The window segmentation is almost twice more performant using polarimetry model than RGB model. Indeed, these results can be explained by the polarization state of such areas.

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed the introduction of polarimetry to pixel-wise road scenes segmentation field. Since to our knowledge there was no dataset with outdoor scenes captured via polarimetry, we created our own dataset. This dataset being made up of several modalities, the key idea was to have a comparison measure. As polarimetric data require meticulous exploitation, we have developed an augmentation method to preserve the physical properties of this modality. This approach defines the possible transformations and provides the necessary formulas for a rotation or flipping. We then used our augmented dataset as input to the SegNet Network to estimate the results. After comparing the SegNet Polarimetry and the SegNet RGB we can deduce that polarimetry offers a considerable advantage over RGB. Indeed, reflective areas are better detected while maintaining or improving the segmentation performance of other areas. We can conclude that polarimetry can provide a new type of information usefull in many fields such as robotics, computer vision or autonomous cars.

However, there are still some areas for improvement. One area for improvement is the use of a more complex network with deeper and more abstract functionalities. This will then allow the results to be compared between a simple network and a deeper network. The immediate objective of improvement is to use raw polarimetric images to eliminate any preprocessing.

ACKNOWLEDGEMENTS

This work was supported by ANR VIPeR, ANR ICUB. We gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research.

Table 3: Per-class Accuracy and accuracy Difference.

	Sky	Water	Windows	Road	Cars	Building	None	Mean
Polarimetry	75.34 %	75.70 %	82.85 %	77.82 %	71.40 %	87.69 %	78.95 %	78.54 %
RGB	89.57 %	78.61 %	44.50 %	78.45 %	48.48 %	67.84 %	83.4 %	69.83 %
Difference	-14.23%	-3.51 %	38.35 %	-0.63 %	22.92 %	19.85 %	-4.45 %	8.71 %

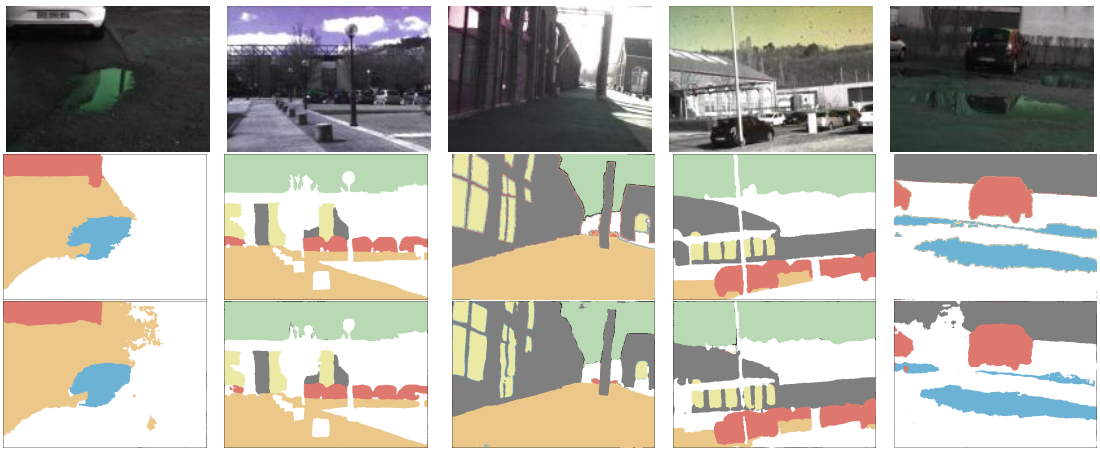


Figure 4: Polarimetry Results - Test Set Output.The top row is the input HSL image. The middle row is the ground truth manually segmented. The bottom row is the prediction output by the SegNet Polarimetry.

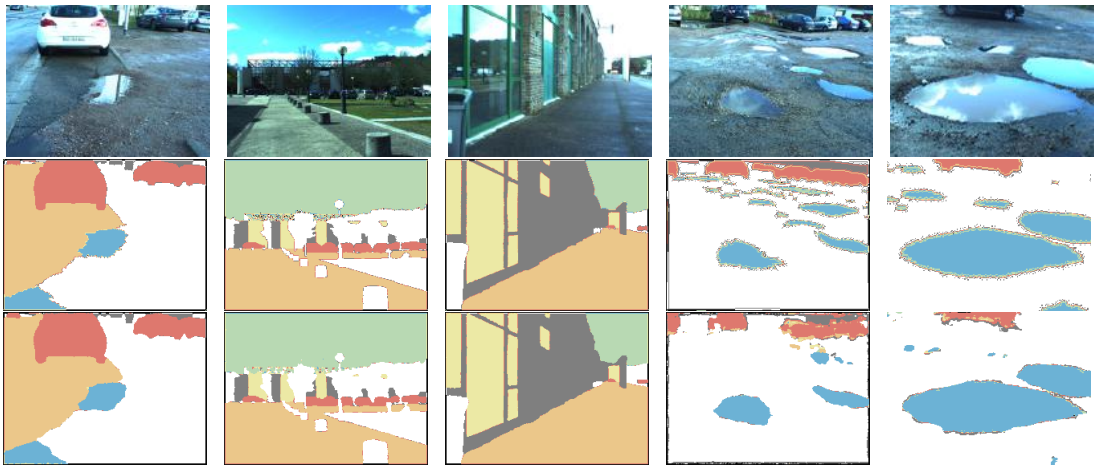


Figure 5: RGB Results - Test Set Output. The top row is the input RGB image. The middle row is the ground truth manually segmented. The bottom row is the prediction output by the SegNet RGB.

REFERENCES

Badrinarayanan, V., Handa, A., and Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *CoRR*, abs/1505.07293.

Berger, K., Voorhies, R., and Matthies, L. H. (2017). Depth from stereo polarization in specular scenes for urban robotics. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 1966–1973. IEEE.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2015). Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2016). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–

848.

Couprie, C., Farabet, C., Najman, L., and LeCun, Y. (2013). Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*.

Gupta, S., Girshick, R., Arbeláez, P., and Malik, J. (2014). Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision*, pages 345–360. Springer.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *arXiv preprint*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Millerd, J., Brock, N., Hayes, J., North-Morris, M., Kimbrough, B., and Wyant, J. (2006). Pixelated phase-mask dynamic interferometers. In *Fringe 2005*, pages 640–647. Springer.

Morel, O., Meriaudeau, F., Stolz, C., and Gorria, P. (2005). Polarization imaging applied to 3d reconstruction of specular metallic surfaces. In *Machine Vision Applications in Industrial Inspection XIII*, volume 5679, pages 178–187. International Society for Optics and Photonics.

Nguyen, C. V., Milford, M., and Mahony, R. (2017). 3d tracking of water hazards with polarized stereo cameras. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 5251–5257. IEEE.

Nordin, G. P., Meier, J. T., Deguzman, P. C., and Jones, M. W. (1999a). Diffractive optical element for stokes vector measurement with a focal plane array. In *Polarization: Measurement, Analysis, and Remote Sensing II*, volume 3754, pages 169–178. International Society for Optics and Photonics.

Nordin, G. P., Meier, J. T., Deguzman, P. C., and Jones, M. W. (1999b). Micropolarizer array for infrared imaging polarimetry. *JOSA A*, 16(5):1168–1174.

Oliveira, G. L., Burgard, W., and Brox, T. (2016). Efficient deep models for monocular road segmentation. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 4885–4891. IEEE.

Pan, S. J., Yang, Q., et al. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 1(2):4.

Rahmann, S. and Canterakis, N. (2001). Reconstruction of specular surfaces using polarization imaging. In *null*, page 149. IEEE.

Rankin, A. and Matthies, L. (2010a). Daytime water detection based on color variation. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 215–221. IEEE.

Rankin, A. L. and Matthies, L. H. (2010b). Passive sensor evaluation for unmanned ground vehicle mud detection. *Journal of Field Robotics*, 27(4):473–490.

Ratliff, B. M., LaCasse, C. F., and Tyo, J. S. (2009). Interpolation strategies for reducing ifov artifacts in microgrid polarimeter imagery. *Optics express*, 17(11):9112–9125.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 242–264. IGI Global.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 5.

Wolff, L. B. and Andreou, A. G. (1995). Polarization camera sensors. *Image and Vision Computing*, 13(6):497–510.

Yan, S. H. (2014). Water body detection using two camera polarized stereo vision. *International Journal of Research in Computer Engineering & Electronics*, 3(3).

Deep learning approach for artefacts correction on photographic films

Strubel David[a], Blanchon Marc [a], Morel Olivier[a] and Fofi David [a]

[a]ImViA Laboratory, ERL CNRS 6000, Universite de Bourgogne Franche Comte (UBFC);

ABSTRACT

The use of photographic films is not totally obsolete, photographers continue to use this technology for quality in terms of aesthetic rendering. A crucial step with films is the digitization step. During the scanning process, dust, scratch and hair (artefacts) are a real problem and greatly affect the quality of the final images. The artefacts correction has become a challenge in order to preserve the quality of these photos. In this article, we present a new method based on deep learning with an encoder-decoder structure to detect and eliminate artefacts. In addition, a dataset has been created to carry out the experiments.

Keywords: artefact removal, film photographic, deep learning, control quality

1. INTRODUCTION

Nowadays photographers and film-makers continue to use analogue films for different reasons, as image quality or out of nostalgia. One of the most crucial steps in the use of analogue film is digitization. Indeed, to exploit the full potential of films, it is necessary to be able to digitize images in high quality in order to use modern post-production tools. When scanning films, scratches, dust and/or hair can affect the quality of the final images. The scanning process is also long and expensive. The standard artefact reduction approaches such as clean-room development or the correction on film are extremely expensive and tedious tasks. In order to overcome this problem, a solution using a convolutional neural network is proposed to restore the images. In this article, we present our dataset[7] and the preliminary results obtained with a SegNet[1] Deep Learning network.

2. STATE OF THE ART

Removing artefact on film has long been a challenge. One of the solutions proposed so far has the disadvantage of dividing the problem in two. Firstly, the segmentation of artefact and on the other hand, in-painting. Richard et al.[6] propose an efficient painting method based on manual selection of imperfection. This method then proceeds to convolutions for in-painting. More recently in Bergnan et al.[2] has proposed a fully automatic solution to detect dust scratches and hair first before removing them. In the work of Bergnan et al.,[2] artefact detection is done locally to provide good quality segmentation of imperfections by the pixel labelling method. Besserer et al.[3] also offers film detection and correction by focusing on the vertical stripes (scratch) commonly observed on

(a) Input (b) Ground truth (c) Prediction

Figure 1: A (c) corrected image obtained for a consequent quantity of artefacts on the (a) input image and comparison with the (c) ground truth.

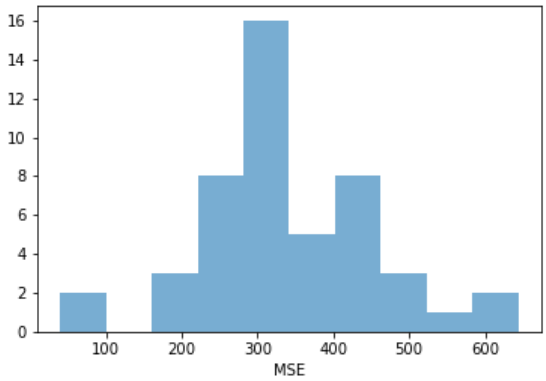


Figure 2: Distribution of MSE for the test-set.

photographic or video films. Due to a mechanical defect in the cameras, the films may be scratched several times but always vertically. The solution provided is based on vertical scraping in order to easily detect and repair it. A similar problem of artefact on film is the digital sensors dust, as in the Zhou et al.[4] method or Dirik et al.[5] The more interesting part in these articles is the segmentation of the sensors dust.

3. RESULT AND TALK

Our hypothesis is to claim the potential effectiveness of convolution networks for the artefacts correction. The method is effective for the two main problems which are the segmentation of artefact and in-painting. A SegNet has been used because of this segmentation capability and its ease of implementation. A quick training (only 100 epochs) was done to evaluate our hypothesis. The result is proposed on 76 images that are not used for training. To evaluate the quality of the results, a MSE (Mean Square Error) metric was used. The MSE average recovered from the test-set is 338.92. The image presented in Figure 1 is then representative of the average with 244.62 imperfections despite a high number of artefacts. The minimum imperfection rate of the test-set is 643.51 and the max is 39.92.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research.

REFERENCES

[1] BADRINARAYANAN, Vijay, KENDALL, Alex, et CIPOLLA, Roberto. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561, 2015.

[2] BERGMAN, Ruth, MAURER, Ron, NACHLIELI, Hila, et al. Comprehensive solutions for automatic removal of dust and scratches from images. Journal of Electronic Imaging, 2008, vol. 17, no 1, p. 013010.

[3] Besserer, B., & Thir, C. (2004, May). Detection and tracking scheme for line scratch removal in an image sequence. In European Conference on Computer Vision (pp. 264-275). Springer, Berlin, Heidelberg.

[4] Zhou, C., & Lin, S. (2007, June). Removal of image artifacts due to sensor dust. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on (pp. 1-8). IEEE.

[5] Dirik, A. E., Sencar, H. T., & Memon, N. (2007, April). Source camera identification based on sensor dust characteristics. In Signal Processing Applications for Public Security and Forensics, 2007. SAFE'07. IEEE Workshop on (pp. 1-6). IEEE.

[6] Richard, M. M. O. B. B., & Chang, M. Y. S. (2001, September). Fast digital image inpainting. In Appeared in the Proceedings of the International Conference on Visualization, Imaging and Image Processing (VIIP 2001), Marbella, Spain (pp. 106-107).

[7] https://github.com/dstrub

LOCAL IN VITRO EVALUATION OF THE BIOMECHANICAL PROPERTIES OF THE ASCENDING AORTIC ANEURYSMS

Siyu Lin (1), Marie-Catherine Morgan (1, 2), Alain Lalande (1, 2), Alexandre Cochet (1, 2), Olivier Bouchot (1, 2)

1. ImViA laboratory, University of Burgundy, Dijon, France; 2. University Hospital of Dijon, Dijon, France

Abstract

This study investigates the mechanics properties of aneurysm of the ascending aortas (AsAAs). The regional variation in mechanical response was compared.

Introduction

An aneurysm of the ascending aorta (AsAA) can progressively evolve over years and can involve an aortic dissection and/or rupture. The surgical replacement of the aneurysmal portion by a prosthesis remains the adopted treatment. Currently, surgical recommendations are based on the maximum diameter of the ascending aorta [1]. It is well known that this factor is not reliable [2]. Understanding the biomechanical properties of the aorta could lead to develop biomechanical criteria, and then to be more accurate in predicting the development of aortic aneurysm. The aim of this study is to obtain the local patient specific elastic modulus distribution of the AsAA from biaxial tensile test.

Methods

Pathologic ascending aortic tissue samples (n = 8) were obtained from patients undergoing elective surgical repair of AsAA. The samples were grouped by the type of valve (tricuspid aortic valve (TAV) and bicuspid aortic valve (BAV) (Table.1).

Patient No.	TAV/ BAV	Maximum Diameter of Aorta	Sex	Age
1	TAV	51mm	M	66
2	TAV	57mm	M	58
3	TAV	67mm	M	59
4	TAV	50mm	M	73
5	TAV	53mm	F	81
6	TAV	70mm	F	78
7	BAV	50mm	M	72
8	BAV	57mm	M	52

Table 1: Patient characteristics grouped by tissue type

All the aortic wall samples were partitioned as medial (MED), posterior (POST), lateral (LAT), and anterior (ANT) quadrants （Fig.1). Each AsAA sample was cut in square size (15 mm x 15 mm, n = 71) with marking the circumferential and longitudinal directions.

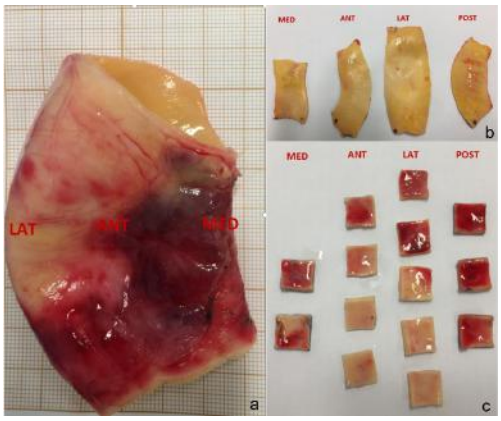


Figure 1: The Pictures of AsAA wall tissue. a) Gross photograph of human AsAA segment; b) Sample pieces from each quadrant (intima view); c) Subsequent processing for obtaining 15mm x 15mm specimens (adventitia view).

For each specimen, an average thickness was measured, using an electronic micrometer (Litematic VL-50, Mitutoyo®). The experiments were carried out by a biaxial tensile test machine (ElectroForce®, Fig 2.).

Figure 2. Pictures of biaxial test machine. a) Picture of LM1 ElectroForce®, TestBench®; b) Picture of sample displacement during the test.

The tissue was under the spray of water (humidity 100%, 32.0 ± 1.0ºC). All specimens were stretched at a rate of 10 mm/min until rupture.

Generally, the aortic wall was assumed to be a non-linear material (Fig.3). In order to calculate the true stress and true strain, we need to define true stress and true strain which depend on the initial dimensions of the tissue. The true stress:

$$\sigma = \frac{F}{A}$$

where A is the current cross-sectional area. The true strain is defined by:

$$d\varepsilon T = \frac{dL}{L}$$

where dL is the instantaneous stretch and L the initial length length of the specimen.

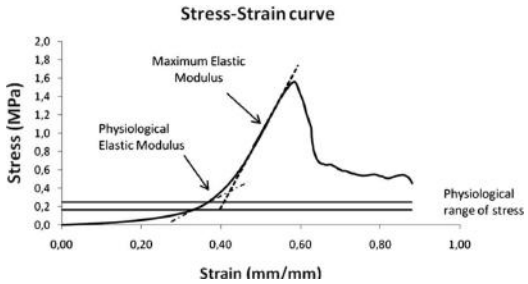


Figure 3. Principles of calculation of the physiological and maximum elastic modulus.

Maximum Elastic Modulus was calculated for each specimen [3].

Results

In the group of TAV (Fig. 4), the lateral wall shows a higher stiffness, roughly the double value of the media.

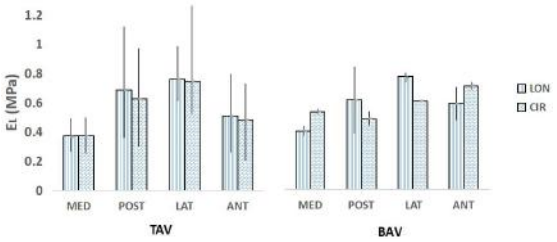


Figure 4: The Maximum Elastic Modulus for the longitudinal and circumferential directions of both TAV (n = 60) and BAV (n = 11) samples for each quadrants.

The thickness of TAV and BAV tissue did not show significant difference in the four regions of the AsAA wall (Fig. 5).

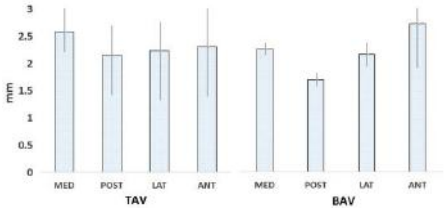


Figure 5: The mean thickness of TAV (n = 60) and BAV (n = 11) for each quadrants.

There was no significant correlation (r = -0.36) between the regional thickness and mechanical properties (i.e. the Maximum Elastic Modulus) of the dilated samples.

Discussion

In this in vitro study we showed a local variability in the aortic stiffness for patients with AsAA without any link with local thickness. More patients would be included to effectively emphasize this regional difference.

References

1. R. Erbel et al. Eur Heart J., 1;35(41):2873-926, 2014
2. A.Duprey et al. Acta Biomaterialia,42:273–285,2010
3. A.Duprey et al. European Journal of Vascular and Endovascular Surgery, 39:700-707, 2010.

Introduction to memristor and applications

Aliyu Isah, Jean-Marie Bilbault
Lab: LE2I
University of Burgundy

***Abstract*—It is part of the goals in this symposium to create awareness and an insight into various fields of research. Here we present an introduction to the fourth circuit element called the Memristor and some of its promising applications.**

***Keywords*—Memristor, Window function, SPICE, Applications**

I. Introduction

IN 1971, Leon Chua [1] posited the existence of the fourth circuit element, the so-called ***memristor*** by observing the symmetrical nature of the three known basic circuit elements; resistor *R*, capacitor *C* and an inductor *L* with respect to the four circuit variables; electric voltage v, electric current i, electric charge q and magnetic *flux* ϕ (see figure 1). The aforementioned three known basic circuit elements linked appropriately any two of the four circuit variables resulting in five possible relationships in conjunction with the conventional definitions of an electric flux and charge. In figure 1, for the sake of completeness there should be six possible relationships, an element is missing that relate magnetic flux and an electric charge. Memristor (*M*) is the contraction of memory resistor, this name is chosen due the fact that memristor remember its previous history hence the memory effect and is non-linear in nature. Depending on the excitation, the device could be charge controlled or flux controlled preferably described by the terms memristance and memductance having units ohms and siemens, equations (1.7) and (1.8) respectively. For more than three decades memristor remained a mystery until in 2008 HP lab announced the realization of memristor in device form [2]. This recent discovery of HP lab attracted many scientists, engineers and researchers to explore more feasible applications of memristor and its device technology. Memristor have versatile applications in discrete and crossbar array configurations.
Since the invention from [2], many memristor device technology emerged that follows similar principle. For example, KNOWM memristor uses electropositive metal (Ag, Cr or Sn) as means of conduction where as HP memristor uses oxygen vacancies that make the device geometry to behave as a semiconductor material. Some other memristor technology are; Purely electronic effects e.g. Ferro electric tunnel junctions (FTJs) memristor, Phase-change memory (Specifically for memory applications), Organic elements e.t.c. It is clear that memristor have a lots of hidden features yet to discover. The three fingerprints of memristor [3] are not enough to characterize an ideal memristor.

$$\text{Flux}(\phi)\ and\ \text{Voltage}(v):\ \phi(t) = \int_0^t v(\tau)dt(\tau) \qquad (1.1)$$

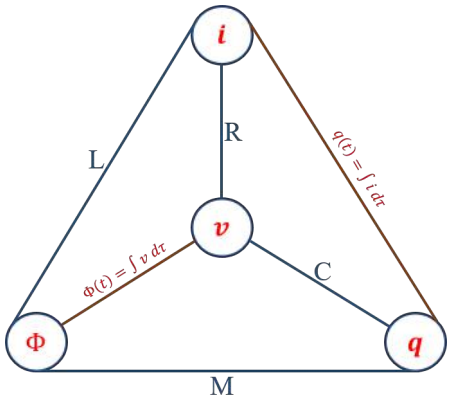


Fig. 1: Circuit elements and circuit variables

$$\text{Charge}(q)\ and\ \text{Current}(i):\ q(t) = \int_0^t i(\tau)dt(\tau) \qquad (1.2)$$

$$f(v,i) = 0, \text{Resistor}(R):\ \mathrm{d}v = R\ \mathrm{d}i \qquad (1.3)$$

$$f(v,q) = 0, \text{Capacitor}(C):\ \mathrm{d}q = C\ \mathrm{d}v \qquad (1.4)$$

$$f(\phi,i) = 0, \text{Inductor}(L):\ \mathrm{d}\phi = L\ \mathrm{d}i \qquad (1.5)$$

$$f(\phi,q) = 0, \text{Memristor}(M):\ \mathrm{d}\phi = M(q)\ \mathrm{d}q \qquad (1.6)$$

$$v(t) = R_m.i(t) \qquad (1.7)$$

$$i(t) = G_m.vt \qquad (1.8)$$

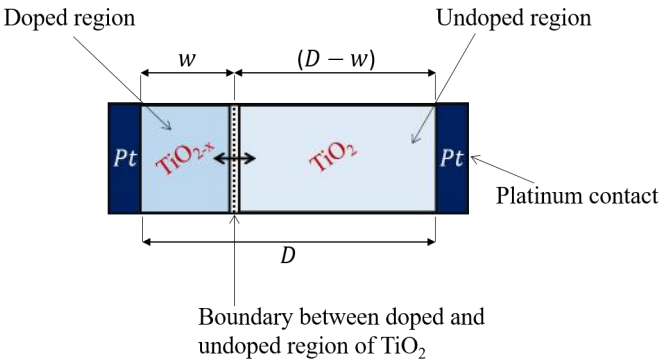


Fig. 2: HP memristor structure

In fig.2, experimentally, memristor is described by state dependent Ohm's law relation, (1.9) and (1.10).

$$v(t) = [R_{on}x(t) + (1 - x(t))R_{off}]\, i(t) \tag{1.9}$$

$$\frac{dx(t)}{dt} = k.f(x,t).i(t) \tag{1.10}$$

Where; $f(x)$ is the window function, $k = \frac{\mu_v R_{on}}{D^2}$ with ionic mobility μ_v, R_{on} is the resistance correspond to the doped region (TiO_{2-x}) and R_{off} is the resistance correspond to undoped region TiO_2. The port equation of the memristor (1.9) is represented by (1.7) and (1.8) according to the type of excitation (current and voltage respectively).

II. Window function

HP memristor suffer an undesirable phenomena called the linear ion drift. The non-linearity drift of the ionic transport tends to seize (hence become linear) as the boundary between doped and undoped region tends toward either of the device edges. Off course there is nothing special when the device behave linearly and it will be just like normal resistor. The window function $f(x)$ (added to the left handside of equation (1.10)) is use to model this phenomena, $f(x)$ not only ensure nonlinear ionic drift but also ensure that the device is operating within the desired boundary limit$\{0$ and 1 or equivalently w and $D\}$. There are many suggested window functions in addition to the one found in [2].Here are some of the suggested window functions;

$$\begin{cases} \text{Strukov et al. HP :} & x(1-x) \\ \text{Joglekar :} & 1-(2x-1)^{2p} \\ \text{Biolek et al. :} & 1-(x-stp(-i))^{2p} \\ \text{Prodromakis :} & 1-[(x-0.5)^2+0.75]^p \end{cases}$$

Using the HP window function and value of parameters for $\mu_v = 10^{-14}m^2s^{-1}V^{-1}$, $R_{on} = 100\Omega$ and $R_{off} = 16K\Omega$, $R_{init} = 11K\Omega$, we obtained the pinched hysteresis voltage-current relationship, satisfied the three fingerprints of memristor.

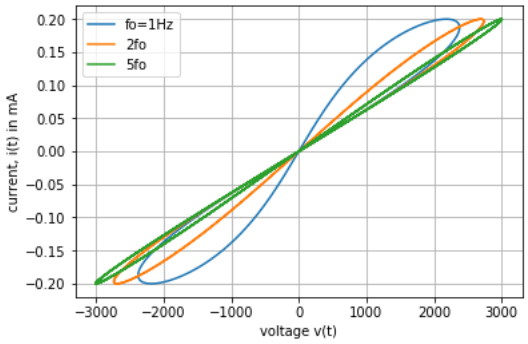


Fig. 3: Current-Voltage plane of a typical memristor at different frequencies

III. SPICE model of memristor

The HP models of memristor is use for modeling memristor in SPICE for researches and simulation purposes. An example of memristor SPICE model is in [4], whose SPICE netlist file is given in Table I. The SPICE netlist file in table I is used as memristor component in SPICE software and the current versus voltage plot is given in figure 4.

```
****** Memristor SPICE model ******
.SUBCKT memristor Plus Minus PARAMS: Ron=100
+Roff=100K Rinit=10K D=10N uv=10F p=1
* STATE EQUATION MODELING *
Gx 0 x value= I(Emem)*uv*Ron/D^2*f(V(x))
Cx x 0 1 IC=(Roff-Rinit)/(Roff-Ron)
Rx x 0 1G
* RESISTIVE PORT MODELING *
Emem plus aux value=-I(Emem)*V(x)*(Roff-Ron)
Roff aux minus Roff
* WINDOW FUNCTION MODELING *
.func f(x)=1-(2*x-1)^2*p ; Joglekar window function
.ENDS memristor
```

TABLE I: Memristor SPICE table

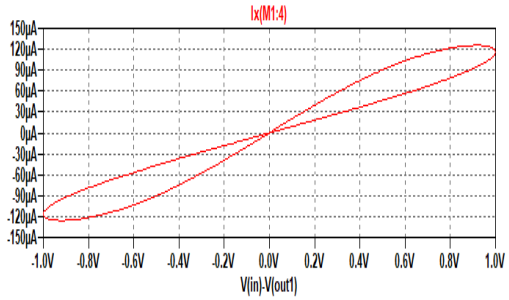


Fig. 4: V-I plot of Table I for v=1V, f=1Hz

IV. Memristor applications

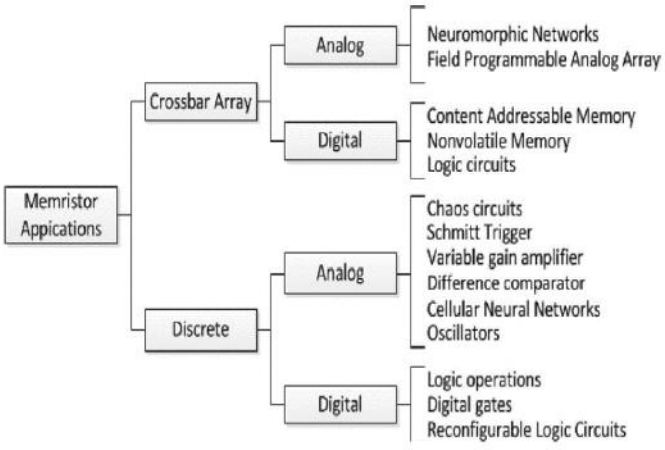


TABLE II: Descrete and array memristor application

V. Conclusion

The emergence of memristor led to the discovery of another two terminal elements - the so-called Memcapacitor and Meminductor. These two elements are not yet realize in device form but their principles theory exist. This put memristor in question as whether is indeed the fourth circuit element, because we can have six elements instead of four and these new three elements looked like extension of the previous three elements. However, memristor is indeed the fourth circuit element owing to the independent nature of its constitutive relationship [5]. However, the discovery of memristor in [2] received criticism due to the incompatibility of (1.9) with Maxwell's equation [6].
Memristor complemented the the possible mathematical relationships of the four circuit variables and it has interesting applications in the fields of science and technology, hence is considered as fitted replacement of the mighty transistor.

References

[1] L. Chua, "Memristor-the missing circuit element," *IEEE Transactions on circuit theory*, vol. 18, no. 5, pp. 507–519, 1971.
[2] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *nature*, vol. 453, no. 7191, p. 80, 2008.
[3] L. O. Chua and S. M. Kang, "Memristive devices and systems," *Proceedings of the IEEE*, vol. 64, no. 2, pp. 209–223, 1976.
[4] Z. Biolek, D. Biolek, and V. Biolkova, "Spice model of memristor with nonlinear dopant drift." *Radioengineering*, vol. 18, no. 2, 2009.
[5] C. Leon, "Everything you wish to know about memristors but are afraid to ask," *Radioengineering*, vol. 24, no. 2, p. 319, 2015.
[6] S. Vongehr and X. Meng, "The missing memristor has not been found," *Scientific reports*, vol. 5, p. 11657, 2015.

Exploration of deep learning-based multimodal fusion for semantic road scene segmentation

Yifei Zhang, Olivier Morel, Marc Blanchon, Ralph Seulin, Mojdeh Rastgoo and Désiré Sidibé
ImViA Laboratory EA 7535, ERL VIBOT CNRS 6000, Université de Bourgogne Franche-Comté, France
{*Yifei.Zhang*}*@u-bourgogne.fr*

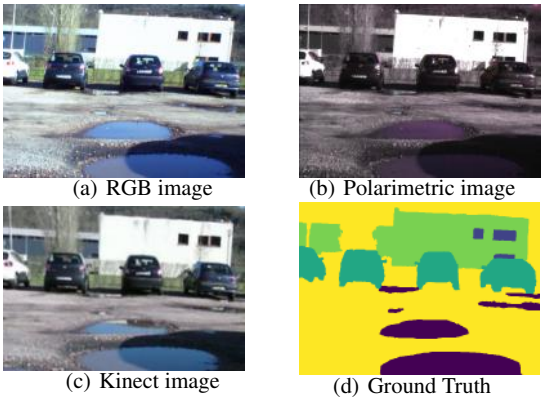
Keywords: Semantic Segmentation, Multimodal Fusion, Deep Learning, Road Scenes

Abstract: Deep neural networks have been frequently used for semantic scene understanding in recent years. Effective and robust segmentation in outdoor scene is prerequisite for safe autonomous navigation of autonomous vehicles. In this paper, our aim is to find the best exploitation of different imaging modalities for road scene segmentation, as opposed to using a single RGB modality. We explore deep learning-based early and later fusion pattern for semantic segmentation, and propose a new multi-level feature fusion network. Given a pair of aligned multimodal images, the network can achieve faster convergence and incorporate more contextual information. In particular, we introduce the first-of-its-kind dataset, which contains aligned raw RGB images and polarimetric images, followed by manually labeled ground truth. The use of polarization cameras is a sensory augmentation that can significantly enhance the capabilities of image understanding, for the detection of highly reflective areas such as glasses and water. Experimental results suggest that our proposed multimodal fusion network outperforms unimodal networks and two typical fusion architectures.

1 INTRODUCTION

Semantic segmentation is one of the main challenges in computer vision. Along with the appearance and development of Deep Convolutional Neural Network (DCNN) (Krizhevsky et al., 2012), the trained model can predict which class each pixel in the input images belongs to. By learning from massive data sets of diverse samples, this method achieves a good performance on end-to-end image recognition. Robust and accurate scene parsing of outdoor environments paves the way towards autonomous navigation and relationship inference. Compared with indoor scenes, off-road perception is more challenging due to dynamic and complex situations. The outdoor environment may easily change in different time slots with light or color variations. Even in structured environments, for instance on urban roads, there are still several challenges such as the detection of glass and muddy puddles.

Most existing datasets and methods for outdoor scene semantic segmentation are mainly based on RGB camera. They are only well acceptable in general conditions excluding complex environment and small amount of samples. To develop additional practical solutions, one of the main challenges is data fusion from multi-modalities. Therefore, considering the RGB modality as a kind of imperfect sensor, we attempt to fuse the complementary feature information of the same scene from other modalities. Actually, several modalities are ubiquitous in robotic systems, such as RGB-D, LIDAR, near infrared sensor, etc. Figure 1 shows the multimodal images of our POLABOT dataset.



(a) RGB image (b) Polarimetric image (c) Kinect image (d) Ground Truth

Figure 1: Multimodal images in POLABOT dataset.

In this work, we use a polarimetric camera, as a complementary modality, to provide a richer description of a scene. Polarization of light radiation has more general physical characteristic than inten-

sity and color (Wolff, 1997). We can figure out that windows of a building, the asphalt road, and the puddle of water have reflected polarizations (Walraven, 1977). Plenty of research have demonstrated that the use of polarization camera can significantly enhance the capabilities of scene understanding, especially for reflective areas (Harchanko and Chenault, 2005).

Over the past few years, a variety of deep learning-based end-to-end approaches have been proposed. One factor that increased the popularity of deep learning is the availability of massive data. In the case without large amount of samples, we attempt to acquire more features of the same scene using several modalities. To some degree, an effective encoding of complementary information enables learning without the need for massive data, therefore the use of small-scale dataset can also lead to good performances. Recent works have shown promising results in extracting and fusing features from complementary modalities at pixel-level. The idea is to separately or jointly train the model using data from different sensors and integrate them into a composite feature at early or late stage.

In this paper, we firstly review the existing fusion methods and datasets in section 2. Next, in section 3, we explore the two typical early and late fusion architectures, and propose our multi-stage Complex Modality network (CMnet), which has an encoder-decoder pattern and takes advantage of the state-of-art segmentation network. We evaluate the performances of the different fusion schemes using two different datasets in section 4. In particular, we introduce a new dataset, which to the best of our knowledge is the first multimodal dataset containing polarimetric images. Finally, the paper ends with concluding remarks in section 5.

2 RELATED WORK

In this section, we go through some of semantic segmentation methods, more details can be found in the review of (Garcia-Garcia et al., 2017). Then we summarize existing deep learning-based fusion schemes and various outdoor scene multimodal datasets.

Deep Neural Network Before deep learning achieved its current tremendous success, traditional computer vision methods were widely used, these methods are base on classifiers which operates on fixed-size feature inputs and a sliding-window. From the beginning with FCN (Long et al., 2015), the end-to-end fully convolutional network has become one of the most popular models for image segmentation. Recent years have witnessed a series of new encoder-decoder architectures along this line, including SegNet (Badrinarayanan et al., 2017), and U-Net (Ronneberger et al., 2015). Followed by the dilated convolutions proposed in (Yu and Koltun, 2015). Based on this technology, the series of DeepLab (Chen et al., 2014; Chen et al., 2018a; Chen et al., 2018b) achieves the state of the art performance in semantic segmentation.

Multimodal Fusion Architecture Benefiting from the improvement of unimodal neural network, excellent progress has been made on multimodal fusion architecture. Several common spectral sensors, such as RGB-D and near-infrared sensor, were applied to pixel-level data fusion of the same scene. For example, FuseNet (Hazirbas et al., 2016) and multi-view neural network (Ma et al., 2017) were proposed to incorporate complementary depth information into RGB segmentation framework. These fusion networks are based on an early fusion architecture, the feature maps from depth are constantly fused into the RGB branch in the encoder part.

Besides, a late fusion based model, Long Short-Term Memorized Context Fusion, also called LSTM-CF, was proposed by (Li et al., 2016). This network extracts multimodal features from depth and photometric data sources separately, then concatenates the feature map at three different scales. Another simple late fusion network (Eitel et al., 2015) was proposed for robust RGB-D object recognition. Furthermore, a convoluted mixture of deep experts technique (Valada et al., 2016a) was used in the late fusion architecture. These early and late fusion architectures were studied and applied to various scenarios and fields, for instance, forested environments navigation (Valada et al., 2016b), urban driving assistance (Jaritz et al., 2018).

Datasets Along with the development of computer vision techniques, a series of high-quality outdoor scene datasets have appeared, such as CamVid (Brostow et al., 2008b; Brostow et al., 2008a), Cityscapes (Cordts et al., 2016), etc. They are widely used in outdoor semantic scene understanding. In addition, some research institutes publish their scenario-based multimodal dataset. For instance, KAIST dataset (Hwang et al., 2015) is a multi-spectral pedestrian dataset of real traffic scenes, which was collected by a co-aligned RGB/Thermal camera, RGB stereo, 3D LiDAR and inertial sensors. Especially for semantic segmentation, there is KITTI dataset (Geiger et al., 2013) which contains high-resolution RGB data, grayscale stereo cameras data, and 3D point cloud; Freiburg Multi-spectral Forest dataset (Valada et al., 2016b) is also a multi-spectral dataset for forested environment semantic

Figure 2: Early fusion and Late fusion architectures comparison.

segmentation, it contains RGB, Depth, NIR, Near-Infrared, Red, Green (NRG), Enhanced Vegetation Index (EVI), and Normalized Difference Vegetation Index (NDVI) images. However, none of these datasets contains polarimetric data.

3 MULTIMODAL FUSION

In this section, we describe the fusion architectures for multi-modalities and the training procedure in details. In essence, the process of training is to minimize the error while regularizing the parameters. Let $S = \{(X_n, y_n)|n = 1,2,...,N\}$ be a set of N training examples, where X_n is the feature vector of n-th example extracted from different modalities, and $y_n \in \{1,2,...,c\}$ is the corresponding segmentation class. Then the training problem can be framed as an optimization one, which can be formulated as:

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L(y_n, f(x_n; \theta)), \qquad (1)$$

where the loss is computed as $L(u,y) = -\sum_k y_k log u_k$. Then we can use, for example, gradient descent algorithm to find local minimum.

3.1 Fusion architectures

In this part, we describe two typical fusion strategies, namely early fusion and late fusion. The two simple structures, as well as their extensions, are widely used for deep learning-based fusion. Here we use SegNet as baseline network to construct such architectures. SegNet has a classical Encoder-Decoder structure followed by a Softmax classifier. The encoder is a regular convolutional neural network which contains five layers. Each layer extracts local features, normalizes the data distribution, obtains sparse representations by means of convolution, batch normalization and ReLU accordingly. Afterwards, pooling is used for downsampling the feature map and propagate spacial invariant features. Correspondingly, the decoder unsamples the shrunk feature map and recover the lost spatial information to full-sized segmentation.

3.1.1 Early fusion

As shown in Figure 2(a), the early fusion architecture has a unitary neural network, fusion takes place before passing into the encoder. Assume that both inputs (for example one RGB image and one polarimetric image) have size 3×H×W, then fused frame will be 6×H×W. So we also call this sort of fusion architecture as channel fusion.

This fusion architecture, combining features before training, seems simple and light. However, it is also more likely to over-fit. To see why, let consider the model's complexity. Let H be a family of functions taking values in $\{-1,+1\}$ with VC-dimensions d_{vc} (Vapnik, 1998). Then, for any $\delta > 0$ and all $h \in H$, the VC-dimension bound (Mohri et al., 2012) can be derived with a high probability:

$$E_{out}(h) \leq E_{in}(h) + \sqrt{\frac{8}{N} \ln(\frac{4(2N)^{d_{vc}}}{\delta})}, \qquad (2)$$

where E_{out} denotes out-of-sample error, E_{in} denotes in-sample error, and N denotes the data points that the hypothesis space can shatter the set. As the amount of input's dimensions increases, so does the VC-dimensions. Then the model complexity $\Omega(N,H,\delta)$ rises along with the increase of VC-dimensions. As a result, larger data samples should be fed to fit the deep neural model for less in-sample error. In other words, in the case that samples are not huge enough, the model may be easier to over-fit.

3.1.2 Late fusion

Figure 2(b) shows the late fusion architecture which was used in this paper. It has two separated branches of network, with each branch trained to extract features from a special modality. Fusion takes place after a series of downsampling. Assuming that the two feature maps have size 1×H×W, after concatenation, the resulting feature will be 2×H×W. Then a 1×1 convolution is applied to reduce the number of channels.

This approach has the advantages that each network computes weights separately while encoding. Compared with early fusion, to some extent, it may

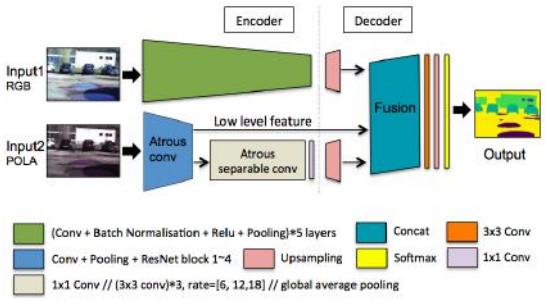


Figure 3: Our proposed fusion architecture: CMnet for multimodal fusion based on late fusion architecture.

reduce the difficulty of model fitting and yield a better outcomes. Furthermore, thanks to the scalability and flexibility of this architecture, the model can be designed in accordance with requirements and easily extend to multi-inputs without a large dimension increase.

3.2 Proposed fusion model

We propose a new approach for multimodal data fusion, Complex Modality Neural Network (CMnet), based on late fusion architecture since it has aforementioned merits.

Let $S = \{(X_n, y_n)|n = 1, 2, ..., N\}$ denotes the training set, and $X_n = \{x_a, x_b\}$ is the training example, where x_a and x_b are the vector of input images from modality a and b, respectively. Also let M_1, and M_2, denote the map between the input and output of the first, and second branch of the encoder-decoder network, respectively. Then the output of the fusion module can be written as:

$$\hat{y}_n = f(X_n) = \text{softmax}[W * (M_1(x_a) + M_2(x_b))], \quad (3)$$

where, W is a series of convolution kernels for upsampling. The *softmax* function is introduced to represent the categorical distribution, and is defined as:

$$\text{softmax}(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, \quad (4)$$

where $z = [z_1, \ldots, z_K]^T$.

Figure 3 presents the whole architecture of CMnet. It has an Encoder-Decoder structure and two separated branches. The encoder is used for mapping raw inputs to feature representations. The decoder integrates three feature maps, then recovers the feature representation to final segmentation results. That is a reliable method to extract different modality features and recover sharp object boundaries for end-to-end segmentation.

On the one hand, the branch for RGB modality incorporates a SegNet-like encoder. By copying the indices from max-pooling, it can capture and store boundary information in the encoder feature maps before sub-sampling. We keep this strength to make the network more memory efficient and improve boundary delineation. On the other hand, we focus on the feature quality of the extra modality. Other modalities can provide rich complementary information on low level appearance features.

However, how to captures rich contextual information from extra modality is a challenging task. We refer to the state-of-the-art segmentation network Deeplab v3+ (Chen et al., 2018b), which uses a new pooling method named ASPP (Atrous Spatial Pyramid Pooling) to incorporate the multi-scale contextual information. We apply this network structure as the other branch's encoder for the complementary modality. The first upsampling stage is subsequently applied to each branch to recover the feature representation to the same fusion size, then we fuse these three feature maps, which contains high-level and low-level multimodal features information simultaneously. The second upsampling stage and softmax are applied to the fused feature map, which produces the final results.

4 EXPERIMENTAL RESULTS

In this section, we evaluate the different fusion models, and report a series of results on two datasets. One is the publicly available Freiburg multispectral forest dataset (Valada et al., 2016b), and the second one is a new multimodal dataset containing polarimetric and RGB data, called POLABOT dataset. In this work, all the networks are implemented based on Pytorch framework with a Nvidia Titan Xp graphics processing unit (GPU) acceleration. The input data was randomly shuffled after each epoch. We initialize the learning rate as 0.0001 and use the contraction segments of pre-trained VGG-16 model and ResNet-101 as encoders. Then we fine-tuned the weights of the decoders until convergence.

4.1 POLABOT dataset

As shown in Figure 4, we collected multimodal images using a mobile robot platform equipped with four cameras: the RGB camera (IDS Ucam), a polarimetric camera (PolarCam), a depth camera (Kinect 2.0), and a near-infrared camera. Our raw dataset contains over 700 multi-modalities images. All the images were acquired, synchronized and calibrated using the Robot Operating System (ROS) framework. Our benchmark also contains 175 images with pixel level ground truth annotations which were generated

Figure 4: Mobile robot platform used for the acquisition of the POLABOT dataset. It is equipped with the IDS Ucam, PolarCam, Kinect 2 and a NIR camera.

manually. These images have been dispatched into 8 classes: unlabeled, sky, water, windows, road, car, buildings and others. Benefiting from the use of a polarimetric camera, our mobile robot platform is more capable of discerning on windows, water and other reflective areas. That allows us to do much more exploratory research on polarimetric images in semantic scene understanding domain. In this paper, we use aligned RGB and polarimetric images as inputs to train the fusion models.

For integrating the acquired images, we apply an automatic homographic method to image alignment (Moisan et al., 2012). This method allows to transform the RGB images with respect to the polarimetric images, and crop to the intersecting regions of interest. Moreover, as deep learning models need large data sets of diverse examples, a certain amount of data should be guaranteed. For this reason, we employ geometric data augmentations to increase the effective number of training samples, including rotation and flipping. Data augmentation and multimodal data fusion help to train deep neural networks on small scale datasets.

4.2 Experimental evaluation

4.2.1 Freiburg Multispectral Forest dataset

We train the segmentation architectures on the public Freiburg Forest dataset first. This dataset was collected by a modified RGB dashcam with NIR-cut filter in outdoor forested environment. It consists of over 15,000 raw images, and 325 images with pixel level ground truth annotations for 6 classes, which are the sky, trail, grass, vegetation, obstacle and others. In this unstructured forest environment, Enhanced Vegetation Index(EVI) was proposed to improve sensitivity to high biomass regions and vegetation monitoring. It shows stronger capacities on feature representation than NIR in the previous work. To extract more accurate information, here in our case, we select EVI images as the second modality input besides the visible input.

Table 1: Performance of segmentation models on Freiburg Multispectral Forest dataset. EF, LF refer to early fusion and late fusion respectively. We report pixel accuracy (PA), mean accuracy (MA), mean intersection over union (MIoU), frequency weighted IoU (FWIoU) as metric to evaluate the performance.

	PA	MA	MIoU	FWIoU
RGB	92.07	89.56	79.87	86.19
EVI	92.05	88.76	79.66	85.82
EF	91.80	88.02	78.95	85.67
LF	92.26	89.52	80.36	86.34
CMnet	**93.02**	**90.06**	**81.64**	**87.68**

Table 2: Comparison of deep unimodal and multimodal fusion approaches by class. We report MIoU as metric to evaluate the performance.

	Road	Grass	Veg/Tree	Sky
RGB	77.18	73.47	89.78	80.66
EVI	81.55	73.50	88.08	76.39
EF	80.78	74.07	86.90	78.68
LF	**82.27**	75.66	88.54	77.68
CMnet	81.01	**76.55**	**90.64**	**83.25**

We crop the RGB and EVI images as size $3\times256\times256$, and use them as inputs correspondingly. We report several metrics to assess segmentation models: pixel accuracy (PA), mean accuracy (MA), mean intersection over union (MIoU), frequency weighted IoU (FWIoU). They are frequently used in semantic segmentation domain.

The results shown in Table 1 show that segmentation using RGB images yields better results than EVI images on the whole. This shows that RGB images provide better high-level features while training. For fusion architectures, late fusion methods outperform channel fusion method as we analyzed in the previous section. Our network yields around $1\% \sim 2\%$ comprehensive improvements comparing with other methods.

The results in Table 2 demonstrate the evaluations by class. We report the main four classes as Road, Grass, Veg/Tree and Sky. For uni-modality network, we can find that EVI shows good performance on Road and Grass classes, and RGB modality has a significant advantage on Sky class, which is susceptible to lighting changes. Moreover, the fusion architecture outperforms uni-modality scheme by integrating complementary multimodal information. In particular, our CMnet model achieved a remarkable results

Table 3: Segmentation performance on POLABOT dataset

Input	Methods	PA	MA	F1	MIoU
RGB	SegNet	87.76	81.44	87.67	64.79
POLA	SegNet	90.51	84.15	90.77	68.58
RGB	E-Fusion	90.25	85.06	90.64	69.48
+	L-Fusion	90.02	84.28	90.11	68.81
POLA	CMnet	**90.70**	**85.90**	**90.92**	**72.59**

on segmentation comparing with other fusion architectures, espe.

A note about the results is that Freiburg Forest dataset was collected from a series of frames, the scene of these frames are homogenized, the structure of each class in these images doesn't fluctuate a lot. The specialization of certain scenes may also reduce the demand on the number of samples.

Some segmentation results on the Freiburg dataset are shown in Figure 5.

4.2.2 POLABOT dataset

In the following part, we report several experimental results on our POLABOT dataset. The metrics shown in Table 3 correspond to pixel accuracy (PA), mean accuracy (MA), F1 score (F1) and mean intersection over Union (MIoU).

We process the RGB and polarimetric images with size $3\times448\times448$. While training the networks, we experimentally found that stochastic gradient descent (batch size=1) doesn't work well. It is reasonable that online learning adds too much instability to the learning process as the weights widely vary with each batch, especially for small scale dataset with multi-classes. As a complement of previous analysis of training on small scale dataset, the data augmentation technology applied to POLABOT dataset gives the additional guarantee for weights learning. As a result, we can find that polarimetric images in our dataset provide high quality feature information, it is a beneficial premise for further data fusion. The overall best performance in this dataset was obtained with CMnet integrating RGB and polarimetric inputs, achieving a mean IoU of 72.59%. It yields around 3% comprehensive improvements comparing with the second best methods.

Some segmentation results on the POLABOT dataset are shown in Figure 6.

5 CONCLUSIONS

In this paper, we explored the typical early fusion and late fusion architectures that extract features from multi-modalities, and extensively evaluated theirs merits and deficiencies. We also proposed an extensible multi-level fusion scheme for semantic segmentation, which adopts advanced deep neural network techniques. It provides design choices for future research directions. We presented comprehensive quantitative evaluations of multimodal fusion on two datasets. The results show the benefits of fusing multimodal features to achieve state-of-the-art segmentation performance on small scale datasets. In addition, we introduced a first-of-a-kind outdoor scene segmentation dataset for road scene navigation, which contains high-quality aligned polarimetric images. We empirically demonstrate that the use of polarization camera enhance the capabilities of scene understanding.

Future work concerns deeper analysis of multimodal fusion network, since there is still plenty room for greater precision. One direction is to add the weights for each input while integrating. Moreover, it is possible to optimize the fusion pattern based on the physical properties of modalities and real-world scenarios.

ACKNOWLEDGEMENTS

This work was supported by the French Agence Nationale de la Recherche(ANR), under grant ANR-15-CE22-0009 (project VIPeR), as well as a hardware grant from NVIDIA.

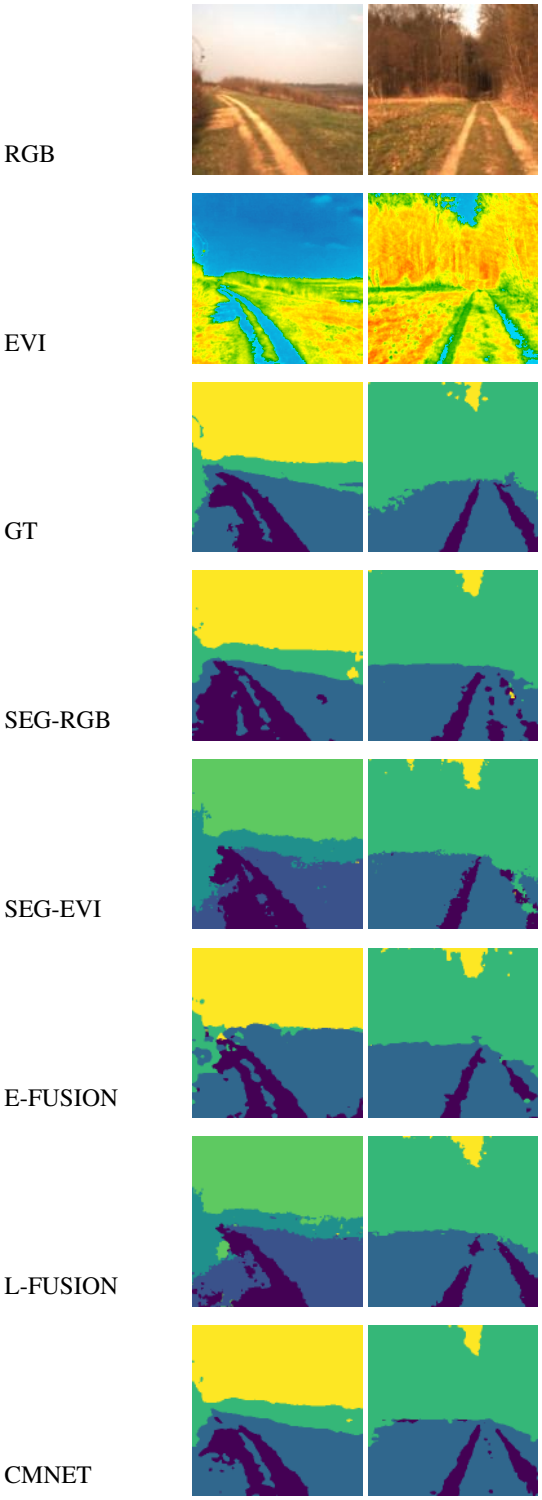


Figure 5: Two segmented examples from Freiburg Forest dataset. RGB and/or EVI images were given as inputs.

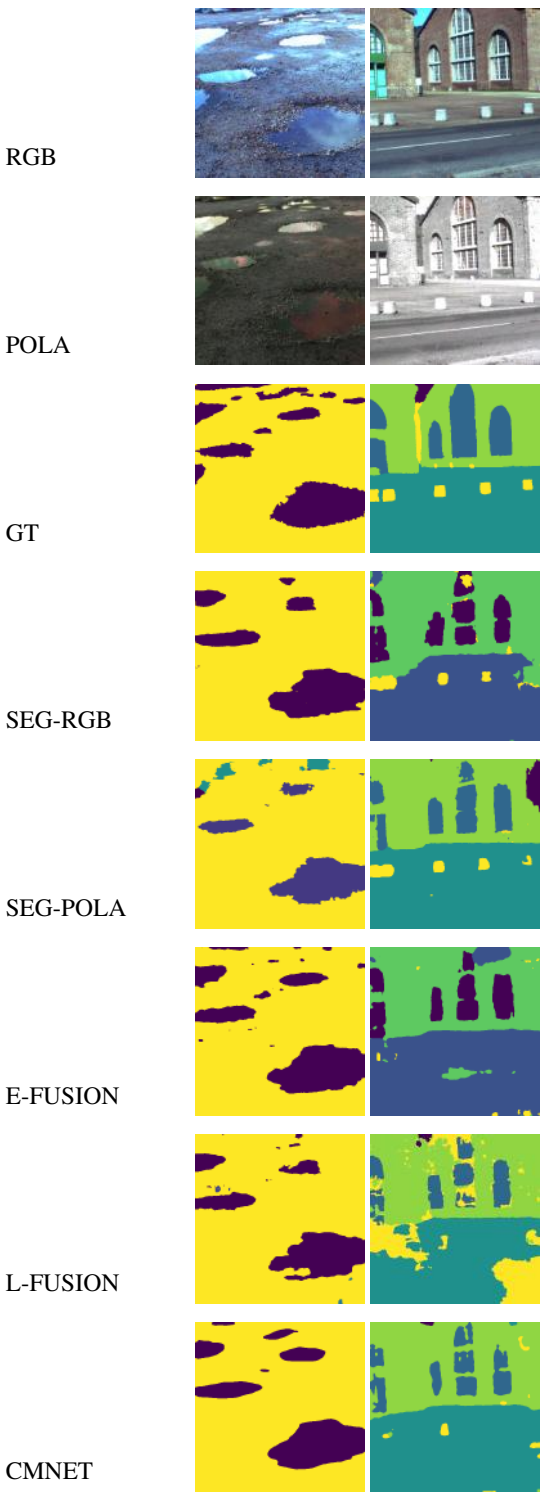


Figure 6: Two segmented examples from POLABOT dataset. RGB and/or POLA images were given as inputs.

REFERENCES

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):2481–2495.

Brostow, G. J., Fauqueur, J., and Cipolla, R. (2008a). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, xx(x):xx–xx.

Brostow, G. J., Shotton, J., Fauqueur, J., and Cipolla, R. (2008b). Segmentation and recognition using structure from motion point clouds. In *ECCV (1)*, pages 44–57.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018a). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018b). Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., and Burgard, W. (2015). Multimodal deep learning for robust rgb-d object recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 681–687. IEEE.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., and Garcia-Rodriguez, J. (2017). A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*.

Harchanko, J. S. and Chenault, D. B. (2005). Water-surface object detection and classification using imaging polarimetry. In *Polarization Science and Remote Sensing II*, volume 5888, page 588815. International Society for Optics and Photonics.

Hazirbas, C., Ma, L., Domokos, C., and Cremers, D. (2016). Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision*, pages 213–228. Springer.

Hwang, S., Park, J., Kim, N., Choi, Y., and So Kweon, I. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1037–1045.

Jaritz, M., De Charette, R., Wirbel, E., Perrotton, X., and Nashashibi, F. (2018). Sparse and dense data with cnns: Depth completion and semantic segmentation. *arXiv preprint arXiv:1808.00769*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Li, Z., Gan, Y., Liang, X., Yu, Y., Cheng, H., and Lin, L. (2016). Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. In *European Conference on Computer Vision*, pages 541–557. Springer.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Ma, L., Stückler, J., Kerl, C., and Cremers, D. (2017). Multi-view deep learning for consistent semantic mapping with rgb-d cameras. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*, pages 598–605. IEEE.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. The MIT Press.

Moisan, L., Moulon, P., and Monasse, P. (2012). Automatic homographic registration of a pair of images, with a contrario elimination of outliers. *Image Processing On Line*, 2:56–73.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Valada, A., Dhall, A., and Burgard, W. (2016a). Convoluted mixture of deep experts for robust semantic segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) Workshop, State Estimation and Terrain Perception for All Terrain Mobile Robots*.

Valada, A., Oliveira, G., Brox, T., and Burgard, W. (2016b). Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In *The 2016 International Symposium on Experimental Robotics (ISER 2016)*, Tokyo, Japan.

Vapnik, V. (1998). *Statistical learning theory. 1998*, volume 3. Wiley, New York.

Walraven, R. (1977). Polarization imagery. In *Optical Polarimetry: Instrumentation and Applications*, volume 112, pages 164–168. International Society for Optics and Photonics.

Wolff, L. B. (1997). Polarization vision: a new sensory approach to image understanding. *Image and Vision computing*, 15(2):81–93.

Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Doctoral day SPIM Vol: 4 (2019)

Geolocalization Ground-to-Aerial using CNN and particle filter

1st Kévin Descharrieres
ERL VIBOT CNRS 6000, ImViA
Universit Bourgogne Franche-Comt
Le Creusot, France
descharrieres.kevin@gmail.com

2nd David Fofi
ERL VIBOT CNRS 6000, ImViA
Universit Bourgogne Franche-Comt
Le Creusot, France
david.fofi@u-bourgogne.fr

***Abstract*—The problem of geo-localization using ground-to-aerial referencement thanks to query images is a very difficult one because of the drastic point of views changing which can often cause matching fails. These last years, many works based on cross-view matching thanks to siamese architecture and convolutional neural networks succeed in this type of geo-localization problems. In this work, we are basing our research on some of these different geo-localization methods and try to improve mainly one of them based on the Siamese architecture [III-A] to do metric learning for the matching task and on particle filter [III-B] to find the different objects and architecture in the environment.**

***Index Terms*—Geolocalization, ground-to-aerial, CNN, particle filter**

I. Introduction

These last decades, the applications which need geo-localization were legion in vision domaines. In the general case, this kind of application is possible if the query image and the reference one are taken from the same point of view like ground-to-ground or UAV-to-UAV in the same area. The simplest case is if both are taken with the least possible changing between both. In our case, the goal is to break these rules and apply the matching between images taken from two different points of view with a different quality of picture and different angles and rotations.

II. Related works

We can classify the different related works of geo-localization in many different methods. In our case, we classify them into three different criterion. The first one is the input dataset [II-A], the second one is the algorithm [III-A] [III-B] and the last one is the environment [III-C].

A. Dataset

The input data set is an important part of geo-localization problems. These data sets are obtain from online services like Google Map, on which one it is possible to obtain different point of views like satellite view or street view. The visual geo-localization became possible thanks to the growing number of accessible photos thanks to online databases. We can observe dataset made by Satellite viewpoints and UAV viewpoints as in "Image registration among UAV Image Sequence and Google satellite image under quality mismatch" by Huang et al. [5] which is working with quality mismatch problems or in "Image localization in satellite imagery with feature-based indexing" by Wu et al. [13] based on a query image of unknown scales and rotation.

Another dataset type is the Ground-to-UAV one. UAV point of view, sometimes called Bird-Eye-View (BEV) is an image taken from an aerial oblic direction. Thanks to this, we can have access to other informations compared to SAT point of view like the facades of the buildings wich can be very useful in geo-localization for matching with street-views. Tian et al. in "Cross-view image matching for geo-localization in urban environment" [11] propose a method for problem of cross-view geo-localization. They used a database of geo-tagged BEV images to find the GPS location of a query image by the building of deep Convolutional Neural Network. Many other works are using these kind of dataset in geolocalization with powerfull results [8], [7].

The last dataset approach we can cite is the Ground-to-Satellite one. This is the case for a big part of the known methods in geo-localization. They are called geo-localization using ground-to-aerial matching. "Accurate Image Localization Based on Google Maps Streeet View" by Amir Roshan Zamir et al. [9] uses a 100,000 images build from Google Maps street view as reference to find the exact GPS location of a query image with a precision of a real GPS.

Some algorithms are already powerful in geolocalization like point cloud algorithms [1] or cross-vew matching [4], [12] using the siamese architecture to find similarities between images [3]. Then, many methods were created thanks to this based on Convolutional Neural Network [4], [2].

III. Our approach

We decribe in this section the approach we are supposing to have to manage this work about geo-localization in view of improving ground-to aerial geo-localization by cross-view matching and possibly perform it accross seasons and environment drastic modifications.

A. Cros-view Matching

In our approach, we are using CVM-Net [2] by Hu et al. [4] as basics. This method is based mainly on the Siamese architecture to do metric learning as said previously. The

first step is to use the fully convolutional layer [1] to extract local image features and then, thanks to the NetVLAD, these features are encoded into global image descriptor.

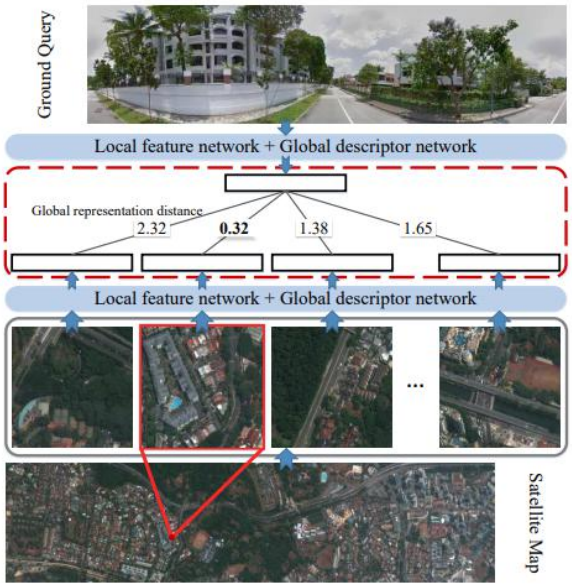


Fig. 1. Illustration of the framework of the CVM-Net [4]

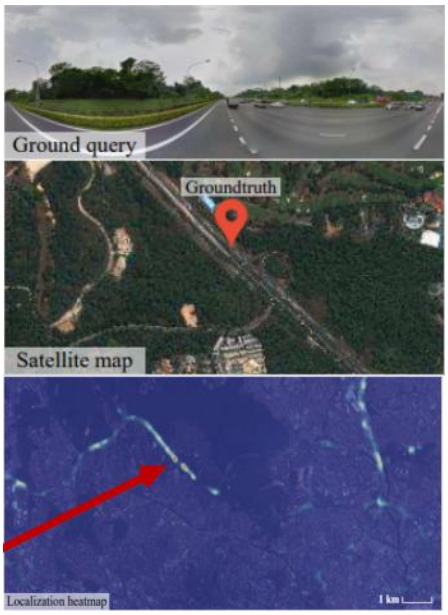


Fig. 2. Example of the matching of the CVM-Net [4]

With this method, we obtained with the first training set (CVM-I) 92.4% of accuracy with a data set from USA and 68% of accuracy with the second network but with the same dataset. At this time, the problem is we want to obtain more precise results in term of localization. We have here only the coresponding area of the street-view location (circa 30m by 30m) and we want an area close to 1m by 1m and the orientation of the streetview view point [III-B].

B. Particle filter

The particle filter approach was developped by Kim and Walter in Satellite Image-Based Localization via Learned Embedings [6]. The goal is to maintain a distribution over the vehicle's pose using this function as an observation model in a filtering network.

This method takes as input a stream of ground-level images and maintains a distribution over the vehicles pose by comparing these images to a database of satellite images [3].

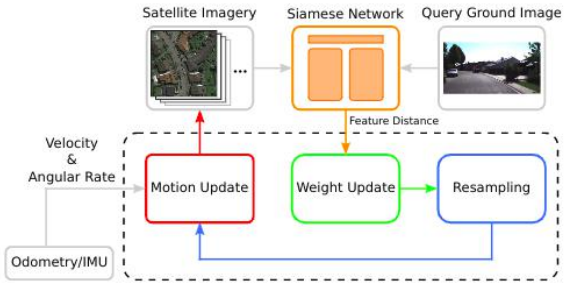


Fig. 3. Global approach of the method [6].

A visualization of the network architecture that consists of two independent CNNs that take as input ground-level and satellite images. Each CNN is an adaptation of VGG-16 CNNs in which mid-level conv4-1 features are downsampled and combined with the output of the last max-pooling layer as the high-level features via summation. The resulting outputs are then used as a measure of distance between ground-level and satellite views [4].

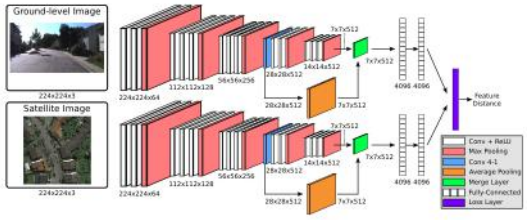


Fig. 4. Convolutional Neural Network used in this method [6].

We will obtain a resulte close the the one in 5 in which one we have a small area and a direction of the street-view camera.

Fig. 5. Example of result we can obtain with particle filter approach [6].

C. Accross seasons and light

A potential last approach of this project is to adapt the CVM-Net method to permit the matching whatever the season, the luminosity of the environment of the query image and the images from the training dataset. To do this, we will base our researches on the method of Sunderhauf et al. called Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free [10] [6]. This method uses Convolutional Neural Network (CNN) to identify matching landmark between images to perform place recognition with extrem point of views and appearances variations. Nevertheless, it doesn't need any form of training because each components are generic enough to be used off-the-shelf.

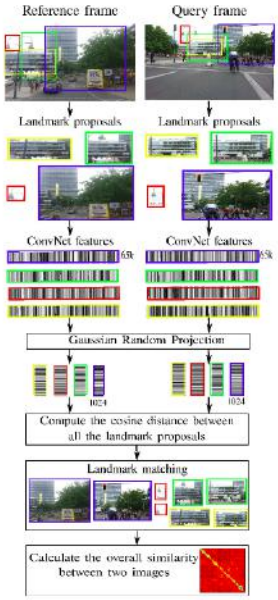


Fig. 6. Example of the matching with environment variation [10]

IV. Conclusion

In this paper, we proposed an idea to developpe a method of geolocalization using cross-view matching with two different point of views accross seasons and with lighting disturbances and motion blur. All these methods use an approch of CNN but the algorithm to manage the environment changing doesn't need a large dataset as input.

References

[1] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M Seitz, and Richard Szeliski. Building Rome in a Day. Technical report.

[2] Relja Arandjelovi, Petr Gronat INRIA, and Josef Sivic INRIA. NetVLAD: CNN architecture for weakly supervised place recognition Tomas Pajdla CTU in Prague . Technical report.

[3] Jane Bromley, Isabelle Guyon, Yann Lecun, Eduard Sickinger, Roopak Shah, Att Bell, and Laboratories Holmdel. Signature Verification using a "Siamese" Time Delay Neural Network. Technical report.

[4] Sixing Hu, Mengdan Feng, Rang M H Nguyen Gim, and Hee Lee. CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization. Technical report, 2018.

[5] Shih Ming Huang, Ching Chun Huang, and Cheng Chuan Chou. Image registration among UAV image sequence and Google satellite image under quality mismatch. In *2012 12th International Conference on ITS Telecommunications, ITST 2012*, 2012.

[6] Dong Ki Kim and Matthew R. Walter. Satellite image-based localization via learned embeddings. In *Proceedings - IEEE International Conference on Robotics and Automation*, 2017.

[7] Tsung Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.

[8] Andras L. Majdik, Yves Albers-Schoenberg, and Davide Scaramuzza. MAV urban localization from Google street view data. In *IEEE International Conference on Intelligent Robots and Systems*, 2013.

[9] Amir Roshan Zamir and Mubarak Shah. Accurate Image Localization Based on Google Maps Street View 1. Technical report.

[10] Niko Sünderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford. Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free. Technical report, 2015.

[11] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[12] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-Area Image Geolocalization with Aerial Reference Imagery. Technical report.

[13] Changchang Wu, Friedrich Fraundorfer, Jan-Michael Frahm, Jack Snoeyink, and Marc Pollefeys. IMAGE LOCALIZATION IN SATELLITE IMAGERY WITH FEATURE-BASED INDEXING. Technical report.

Use of polarimetric imaging to improve the quality control of welding: Detection of oxides

ABIR ZANZOURI KECHICHE,[1*] OLIVIER AUBRETON,[1] ALEXANDRE MATHIEU,[2] ANTOINE MANNUCCI[2],CHRISTOPHE STOLZ [1]

[1] LABORATOIRE ERL VIBOT, CNRS 6000,ImViA, UNIV. BOURGOGNE FRANCHE-COMTE, 12 RUE DE LA FONDERIE, LE CREUSOT 71200, FRANCE
[2]LABORATOIRE ICB, UMR CNRS 6303, UNIV. BOURGOGNE FRANCHE-COMTE, 12 RUE DE LA FONDERIE, LE CREUSOT 71200, FRANCE

*Corresponding author : abir.kechiche@u-bourgogne.fr

Received XX Month XXXX; revised XX Month, XXXX; accepted XX Month XXXX; posted XX Month XXXX (Doc. ID XXXXX); published XX Month XXXX

This paper presents a contribution of polarimetry to active imaging using polarimetric multi-imager, which addresses several technical challenges associated with quality control in industrial application such as welding process. A polarimetric multi-imager system captures different polarization states of an object simultaneously, allowing us to watch dynamic scenes. In addition, this concept allows us to calculate the polarimetric parameters that characterize the weld pool when the metal is in a liquid state. The use of additional information provided by polarimetric imaging for extensions in near infrared wavelengths helps us to detect presence of floating oxides at the weld pool surface. © 2018 Optical Society of America

Polarimetric imaging, welding process, Metals

http://dx.doi.org/10.1364/OL.99.099999

Current advances lead to a real evolution of polarimetric imaging. Polarimetry has proved efficiency in the context of enhanced vision through turbid media [1], machine vision [2], and industrial quality control [3, 4]. Polarimetry provides additional informations on a scene regardless of its wavelength in visible and in infrared spectrum. For example, topography of the surface, composition of the material and quality of the surface will affect the polarization state of the reflected, scattered or emitted light. Generally, this method exploits the polarimetric properties of a light reflected on a specular or diffuse surface. The specular reflection has a linear polarized component oriented perpendicular to the plane of incidence, depends on zenith angle of the rays [5], and may allow a determination of 3D shape of the surface [6]. For our approach, we rely on the polarimetric properties of the infrared radiation emitted by the weld pool surface [7]. The polarization state of the emitted radiation makes it possible to estimate the 3D shape of the surface of the weld pool, as it has been demonstrated in [8]. We present an imaging polarimetry design that uses a polarimetric mutli-imager, addressing several technical challenges associated with quality control of industrial applications such as welding process. This compact design will allow us to obtain simultaneous images of a moving scene with different rotation of the polarizer filters. In addition, this concept makes it possible to calculate the polarimetric parameters that characterizes a non-polarized mobile scene. In this letter, we show that we can do in-situ industrial process quality control such as welding process in order to highlight floating oxide presence at the surface of a metallic molten pool, all this work allows to have new quality control technique using the additional information provided by polarimetric imaging for extensions in infrared wavelengths.

In this letter, main idea is to use a passive polarimetry approach to improve the quality control of a welding process. Welding artefacts can be detected such as presence of oxides flowing overs a weld pool surface. This one is determined from polarimetric analysis of the near-infrared radiations emitted by a weld pool composed of liquid steel, Chromium and Nickel, i.e.304 alloy. While the thermal emissions have a low signal-to-noise ratio for objects near room temperature [9], this ratio is high for the elevated temperatures encountered in welding. Three (S_0, S_1, S_2) out of the four unknown Stokes parameters of the thermal radiations are determined by interpolating the equation for each corresponding pixel in each sub-image of a linear polarizer at an angle α [10].

$$\begin{cases} I(\alpha) = {}^1/_2 . (S_0 + S_1 \cos 2\alpha + S_2 \sin 2\alpha) \\ I(\alpha) = {}^1/_2 . (1 + \rho \cos(2\alpha - 2\varphi)) \end{cases} \quad (1)$$

To probe the complete polarimetric properties of a light source, one needs to measure the Stokes vector S given by

$$S = \begin{pmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{pmatrix} = \begin{pmatrix} I_x + I_y \\ I_x - I_y \\ I_{+45°} - I_{-45°} \\ I_R - I_L \end{pmatrix}, \quad (2)$$

from which the degree of polarization (DOP) of the source can be obtained using the relation $DOP = \sqrt{S_1^2 + S_2^2 + S_3^2}/S_0$. Hitherto, various techniques have been employed to fully or partially

measure the Stokes vectors of an image and obtain the polarimetric information of the scene of interest.

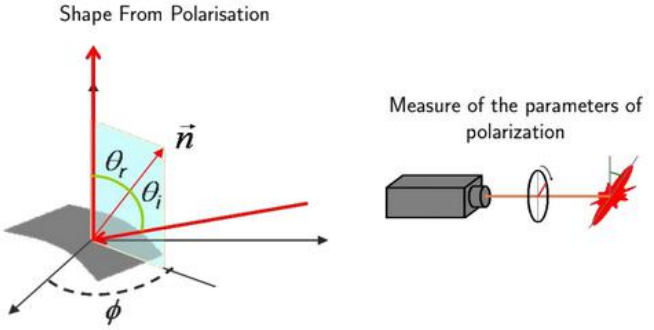


Fig.1.Geometry of the specular reflection on a surface and partial Stokes polarization measure by using a rotating polarizer

The experiment presented here is designed to implement a system using a monochromatic near-infrared polarimeter associated to a stationary Gas Tungsten Arc Welding process. It was performed at 120 A, (Direct Current Electrode Negative), with 4 mm-arc gap on a 200x200x20mm^3 304 block using helium shielding gas. The weld pool is observed with a Phantom V9.1 camera at a grazing angle of 23° through a 810nm narrow bandpass filter during welding (Fig. 2). Nevertheless, the 810 nm wavelength was chosen for observation because of the low absolute quantum efficiency (<0.1) above 810 nm of the camera and the blind spectral window at 810 nm of helium plasma. Fitted on the Phantom V9.1 camera, a wavefront division polarimetric system [11] equipped with four Polarcor linear polarizers 05P109AR.16 from Newport makes, simultaneously, all polarization measurements for every pixel of the dynamic scene (Fig.3).

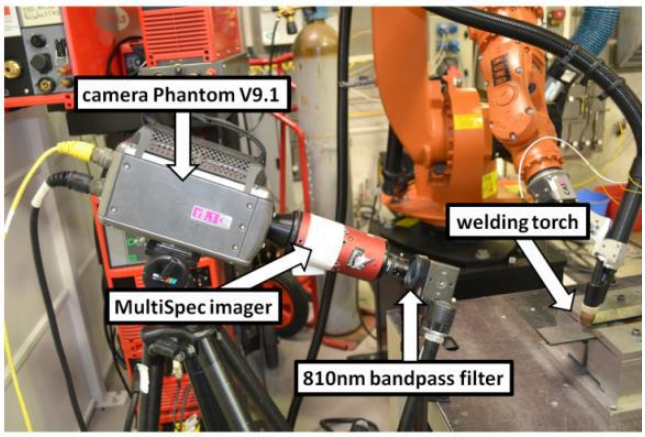


Fig.2. Wavefront division polarimetric system

The polarimetric system captures different polarization states of an object at one single instant, thus suitable for observing dynamic scenes. The simultaneous four-imaging system avoids the problem of time lag between rotation steps and shift in perspective projection of the scene onto the image plane (optical distortion) [12]. In this configuration with linear polarizing filters, the polarimetric imager is a partial Stokes polarimeter that evaluates only three (S_0, S_1, S_2) out of the four Stokes parameters (Fig.1). Low intensity current was preferred to reduce background radiance of arc plasma [13]. Lens diaphragm to f/2.8, corresponding to an approximate aperture diameter of 8.93 mm, and an exposure time of 300 µs are setup. Calibration of the system was required to quantitatively analyze the emission. The relative grey-level attenuation coefficients of each light path (coefficients of 0.75, 1.00, 1.33, and 0.86 for sub-image at 0, 35, 90, and 155 degrees, respectively) were evaluated by acquiring an image without polarized filters and applied on the corresponding polarized sub-image.

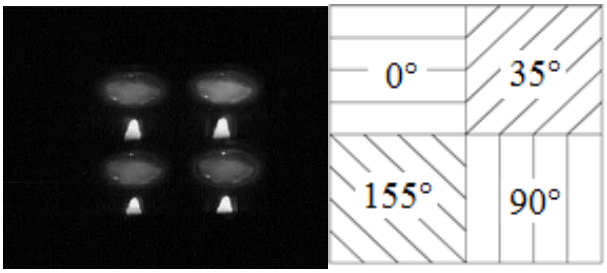


Fig.3. Raw focal-plane image (1200x1600 pixels) showing four polarization channels at indicated orientation.

With our polarimetric system we obtain by calculation the stokes parameters $(S_0, S_1$ and $S_2)$ of the liquid metal of the welding pool scene. These images will serve us as data to obtain an image of the degree of polarization (DOP) which must be correlated pixel by pixel. For that it is necessary to use a feature-based method for structure and motion estimation. It is important to save the problem of the correspondence of every pixel in the image.

These relations are the part of the position filter translates or/and rotates between two views or more views. Then the image matching relationship is the epipolar geometry of the view-pair. A planar homography, also know as a plane projective transformation, or collineation, is specified by eight independent parameters[14]. The homography is represented as a 3x3 matrix that transforms homogeneous image coordinates as:

$$x' = Hx$$

For this paper, an automatic feature-based algorithm allows us to compute a homography between two images using the Internal Point Optimization (IPO). The point feature used and developed by Harris [15], are known as interest points or "corners". We have the algorithm run on an image base with a reference image and a template image in order to find the homography matrix. Once the matrix is calculated. We can test on any image to correct the correlation error between them. All this aims to obtain images of the weld pool with the best possible correlation in order to obtain a polarization degree image with minimum possible error, ie 0.44 pixel.

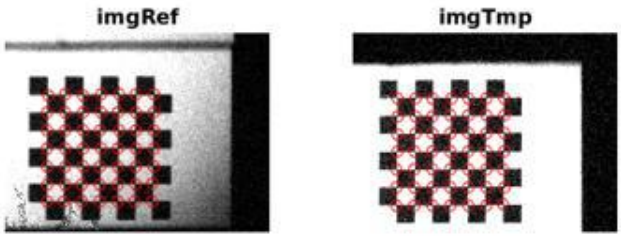


Fig.4 .Reference and template image for homography

Given the inlying interest point correspondences:

$\{x_i \leftrightarrow x_i'\}, i = 1 \dots n$, the final estimation of the homography is obtained by minimizing the following cost function,

$$\sum_i d(x_i, \hat{x}_i)^2 + d(x_i', \hat{x}_i')^2$$

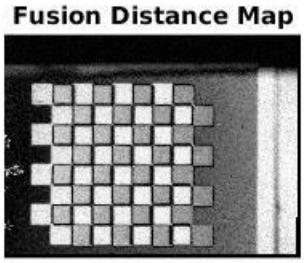


Fig.5.Geometric distance between the images

Where $d(x, y)$ is the geometric distance between the image points x and y. The cost is minimized over the homography $\hat{H}$ and corrected points $\{\hat{x}_i\}$ such that $\hat{x}_i' = \hat{H}\hat{x}_i$. This gives the maximum likelihood estimate of the homography under the assumption of Gaussian measurement noise in the position of the image points.

A metal plate sample was prepared to test the oxide detection during a welding procedure. The type of this metal is 304 grade stainless steel, this is generally considered as the most common austenitic stainless steel. High amounts of chromium and nickel give the 304 stainless steel excellent corrosion resistance.
For this, we did an image acquisition with our experimental system described in the previous paragraph (figure 3). An example of degree of polarization (DOP) is given in figure (6) after registration of the four images. We noticed with the help of the expert: that following the image of the DOP obtained, it is possible to detect, in-situ, the presence of oxides at the pool surface of welding using polarimetric information. For this, we know that the information contained in the image of the DOP essentially comes from the Stokes parameters (S_0, S_1 and S_2), see figures (6) and (7).

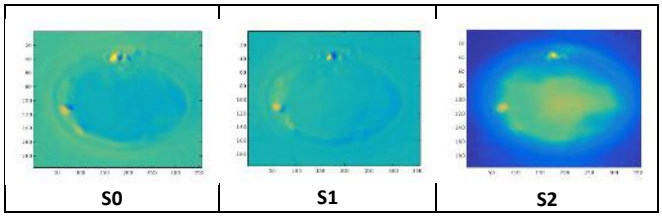


Fig.6 .Image of Stokes parameters: S0 S1 and S2

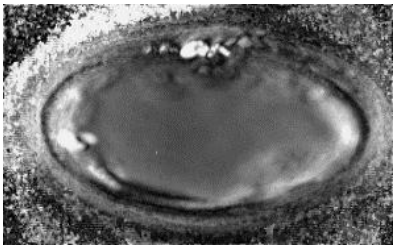


Fig.7 .Image of the degree of polarization"DOP"

We want to make an objective analysis of the information provided mainly by the Stokes parameters S_1 and S_2. A representation of these parameters has been made: a first Gaussian distribution of S1 and another for S_2 according to their probability density. We have obtained a well-known bell-shaped curve when we are on a Gaussian one-dimensional distribution, see figure (8).

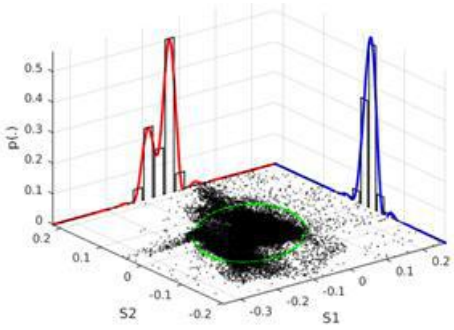


Fig.8 .Gaussian distribution of S_1 and S_2 in function of their probability density

Our ultimate goal is to detect the presence of oxides at the weld pool surface from which we are invited to study the intensity of the connection that can exist between these variables. Following, the representation we can note that our data are correlated, as an ellipse shape. It represents the data of the weld pool without containing the data of the oxidized particles /slags present in the weld pool. Ellipse represents the limits between the data of the welding and the data of the oxidized particles present in the weld pool. What interests us is the cloud of points which is outside the ellipse which represents the position of the oxidized particles inside the weld pool.

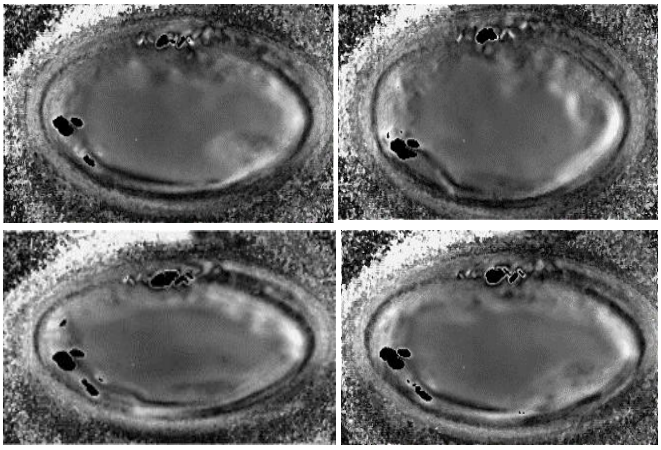


Fig.9 .Image of the degree of polarization (DOP) and image of DOP with black area represent the position of the oxide

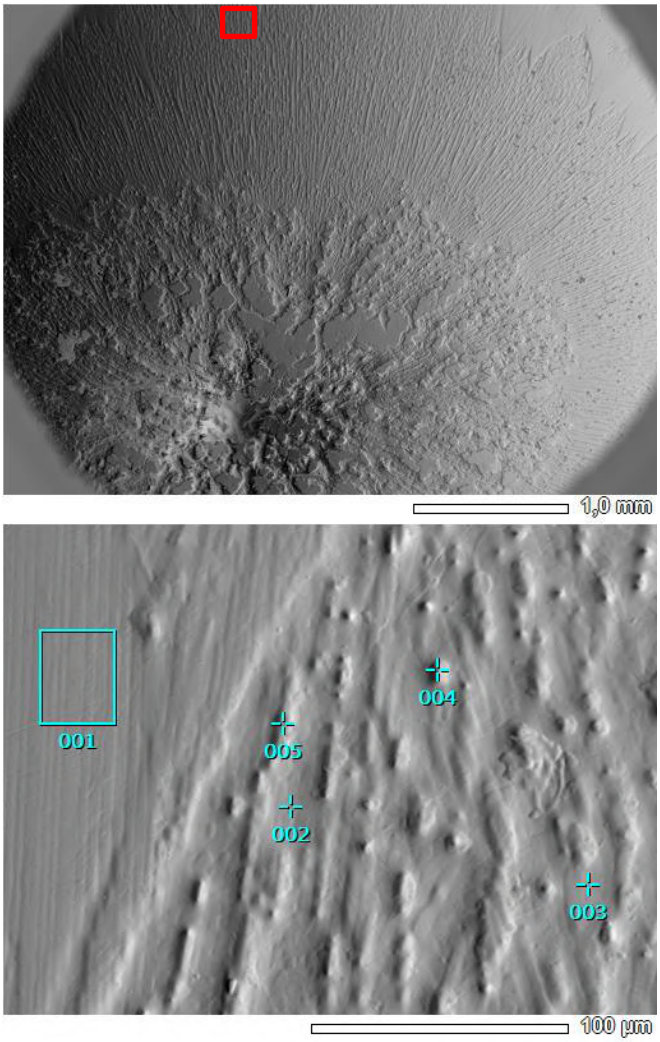


Fig. 10. . Solidified welding pool image with MEB observation

For the analysis of the images, we used the Micrography of MEB images we found areas rich in Oxygen and Aluminum. This is Alumina (Al203), which floats on the weld pool (the black spots detect on the DOP image). Since alumina has a lower density than iron and a much higher melting temperature, it floats on the surface of the pool.

In conclusion, despite the various complications of the welding application, let us mention the use of infrared radiation at wavelengths inside a blind spectral window of the generally very bright arc plasma: the state of polarization provides access to a wealth of information related to the welding process. The polarimetric data can detect the presence of oxides inside the surface of the weld pool in the liquid state. This proves that polarimetry can provide us with essential information to improve the quality control of welding processes. We have detected a weld failure among several, without the need to go through the 3D reconstruction of the weld pool. We want to try to see if we arrive with the polarimetry to find topology information of the weld pool (concave / convex) without going through the 3D reconstruction.

References

[1] S. Demos and R. Alfano,"Optical polarization imaging,"Appl.Opt.36, 150–155 (1997).
[2] M. Yamada, K. Ueda, I. Horiba, and N. Sugie,"Discrimination of the road condition toward understanding of vehicle driving environments," IEEE Trans. Intell. Transp. Syst.2,26–31(2001).
[3] P. Terrier, V. Devlaminck, and J. Charbois,"Segmentation of rough surfaces using a polarization imaging system,"J. Opt.Soc. Am. A25, 423-430 (2008).
[4] O. Morel, C. Stolz, F. Meriaudeau, and P. Gorria,
"Active lighting applied to three-dimensional reconstruction of specular metallic surfaces by polarization imaging,"Appl. Opt.45,4062–4068 (2006).
[5] L.B. Wolff and A.G. Andreou, Image and Vision Computing 13 (6), 497 (1995)
[6] Nicolas Coniglio, Alexandre Mathieu, Olivier Aubreton, and Christophe Stolz ,Characterizing weld pool surfaces from polarization state of thermal emissions,Optics Letters,Vol. 38,Issue 12,pp. 2086-2088(2013)
[7] K.P. Gurton, R. Dahmani, and G. Videen, Measured Degree of Infrared Polarization for a Variety of Thermal Emitting Surfaces (ARL-TR-3240, 2004).
[8] A. Zanzouri Kechiche,R.Rantoson,O. Aubreton, F.Meriaudeau, and C.Stolz,Shape from polarization in the far IR applied to 3D digitization of transparent objects, Qirt 2016 Gdansk Pologne.
[9] D. Miyazaki, M. Saito, Y. Sato, and K. Ikecuhi, J. Opt. Soc. Am. A 19, 687 (2002).
[10] O. Morel, C. Stolz, F. Meriaudeau, and P. Gorria, Appl. Opt. 45, 4062 (2006).
[11] Tyo S.J., Goldstein D.L., Chenault D.B., and Shaw J.A., Appl. Opt. 45, 5453 (2006).
[12]Wolff L.B. and Andreou A.G., Image and Vision Computing 13 (6), 497 (1995).
[13] Weglowski M.S., Journal of Achievements in Materials and Manufacturing Engineering 20 (1-2), 519 (2007).
[14] R.I. Hartley ,In defence of the 8-point algorithm, Proceedings of IEEE International Conference on Computer Vision, DOI: 10.1109/ICCV.1995.466816
[15]C.J.Harris and M.Stephens. A combined corner and edge detector.In Proc.4th Alvey Vision Conference, Manchester, pages 147-151,1988.

Doctoral day SPIM Vol: 4 (2019)

Active thermography, non destructive testing and mobile stereovision system

Thomas Herrmann
Olivier Aubreton, University of Burgundy, ImVia Laboratory CNRS ERL 6000,
12 rue de la Fonderie, 71200, LeCreusot, France
Cyril Mignot, Cyrille Migniot, ImViA EA 7535, Univ. Bourgogne FRanche Comté, Dijon, France.

Abstract

Active thermography testing is not a new technology but is still widely used in industrial areas because it can be used for various applications. However, even if the technology is already quite old it is clear that we can still optimize the process or find other means to use it. For now, the technology is mainly used with statics systems and is not really used on mobile robots or other systems. What we want to do is to create a new autonomous system which can move and take measures of an object by taking multiples 3D points at its surface. This project, based on old works[1][2] still need many problems to be solved. We need to create a two-camera system with the limitations induced by thermal cameras and we need to make use of our thermal marker to follow the move of our robots. We will explain here the pipeline of our project.

1. Introduction

Non destructive testing is required by many companies. As indicated by his name, it consists in checking the integrity of an object without breaking or damaging it. It can be used, for example to check the inside of metal parts without destroying a fully finished item. Non destructive testing can be used with many other tools to get various results depending of what the user want to check, for example, use light, ultrasound or heat. For our project we used the head produced by a laser.[3]

1.1. Thermography

Thermal imaging is used to capture the temperature at the surface of an object. This method use a thermal camera, which is similar to a standard, RGB camera but can capture different wavelength. When associated with non destructive testing, this technology is used to monitor the temperature at the surface of the tested object and to record any variation of those values. Generally, if the tested object do not produce heat, an external heat source is placed near it, this is called active thermography. It can be a heater, a spotlight, a laser, etc. In this case, the heat produced outside of the object will penetrate inside it. The interesting thing is, that the penetration speed will be determined by the state of the
material inside the tested object. If the object is not made of an unique material the heat will be in part reflected when reaching the discontinuity taht will induce a temperature raise on the surface of the object. This will be used to estimate the state of the material inside the object. For example, if an object made of a block of metal without any discontinuity show variation of temperature on its surface, that can be the sign
of the presence of some internal issues. The time when those high-temperature areas will take to appears will be used to calculate the depth of the crack when the size of the heat spot can be used to calculate the size of the affected area.

1.2. Static and dynamic systems

Generally, non destructive testing and thermal imaging are used in a static system. The cameras are statics, the heat source the tested item too. What we want to study here is a mobile system. That's mean we need to get one of those element to move in space, not just in a plane. To do this without adding any outside elements wee need to use another thermal camera to create a stereo-vision system based on thermal camera. This newly added camera will provide us 3D points of the object. To do this we need to work on a 3D representation of the cameras compared to the object.[4]

2. Thermal camera

A thermal camera is calibrated like a normal, RGB camera to compute its intrinsic parameters. However, in this paper case, we need to calibrate one only one camera but two to create a 3D scene representation. To do this we need to link the representation of the first thermal camera to the second one. We are using a chessboard target to compute the translations and rotations matrix between those two cameras.

2.1. Chessboard target

To calibrate a camera we need to use a known object. Generally we take several chessboard images and we try to find the intersections at the corners of the white and black squares. This type of calibration target is used to compute the intrinsic parameters of the camera. To do so, we need to have several lines and columns of squares and at the very least 8 points to calculate the associated matrix for each image.

Figure 1: A classic paper chessboard calibration target

The more the number of point is, the more correct the final matrix will be, however having to much points is not recommended. Generally, we keep the numbers of points between 20 and 100. All the calibrations targets are not with a chessboard pattern. Some of them also have several (as)symmetric circles. When using this type of calibration pattern the numbers of points we can put on the target is generally less but they are also well-used.

2.1. Thermal chessboard

Using a normal paper chessboard calibration target with a thermal camera is not possible. That of course due to the fact that a thermal camera can only see heat, and, basically, a sheet of paper has a constant surface temperature. We need to use another type of material to get reference points for a thermal camera. Generally, a thermal chessboard is made of copper and plastic. Because we sometime need to heat the chessboard directly or with the tested item, the plastic need to be somewhat heat-resistant. Those kinds of calibration targets works well because the copper parts dissipate heat better than plastic. Even when the chessboard is not heated, the copper square will appear “colder” on the thermal camera. However to strengthen the contrast between the two materials it can be a good idea the heat the target because temperature differences will be easier to be see.[5]

Figure 2 : A small copper/Plexiglas calibration target

2.2. Issues with thermal calibrations targets

As we already said it, when using paper calibration targets, we only need to identify the black and white parts on the target and the corner detection isn’t very hard to compute after this. This method is very robust because those two colors are not subject to a lot of variation, even in bad light condition. The black will always stay black, only the white part can change. To separate the black square from the white squares we binarize the image. Of course, this binairization can be adaptive be the result is generally good with paper targets. In normal light condition, on a 8-bits picture, the blacks areas are often less than 20 and the white value over than 220. That’s mean we only need 25% of the color scale to have all the squares, and we have all the left scale to see the difference, which is very good to make a precise picture. However, when working with a thermal picture and with a thermal calibration target, the precision isn’t the same. We often have cold or hot areas outside of the target which will reduce the rift range between “black” and “white”. For a thermal target, we have, for example black under 120 and white after 160, which only left 40 colors between the two final colors. The issue is that when the target is not directly in front of the camera, some copper pixels from a corner of the chessboard can be of the same color than a plastic pixel on the other side.

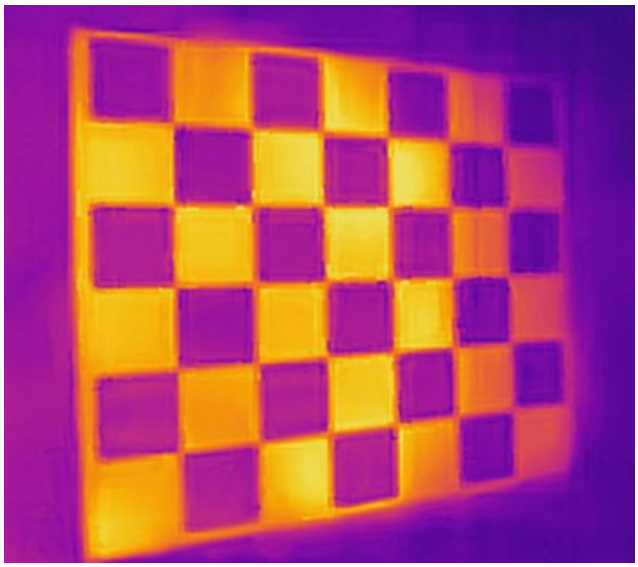


Figure 3: Hard-to-detect chessboard calibration pattern

On the figure 3 we can see that the yellow corners turn slowly to purple the more we go to the right. Sadly, it’s the same purple as the purple square on the leftmost side of the picture. Also, another issue with the corner detection is the picture resolution. If we can easily get a 1920*1080 pixels picture with a color camera, thermal camera usually can’t provide picture with a resolution bigger than 640*480 pixels. Sometimes it’s even smaller than this.

3. Movement tracking

One of the main aspect of this work was to create a mobile system. We have already tell that we need two cameras to compute the 3D coordinates of the object but we have not said that this computation will be the only method we want to use to move the system. What we want exactly do is to use a laser to create points on the surface of the tested item. The point will heat the surface of the object so we will be able to do some acquisition and controls thanks to this, but after a moment, when the surface will be hot enough, we will stop the laser, immediately move the system elsewhere and, just a few seconds later make another set of points with the laser. For a short time we will be able to see the two areas where the laser has heated the surface and we will be able to use those points to compute the movement of the system between those to areas.

3.1. Computation

The computation of the system movement will be done by calculating the distance between the two 3D heat points we have. We can separate the system movement into translation and rotation and follow this movement with some in-between picture if we need to do. A very important thing is, that the movement between to position can't be big because the heat points will disappear fast. The move range is limited by them but also by the tested object's material and the cameras' windows two.

3.2. Heat points

Heat points made with a laser can only stay some seconds, even when the right material is used. Firstly, the mobile system won't be able to be use on metal object for now because tests has shown that metal can dissipate a heat point in less than half a second which isn't enough time to move the system. On wood or plastic, this dissipation is also fast (from 3 to 10 seconds) but in a lot of cases enough to move and make another point. It is important to note that the heat on an old laser hit spot will decrease to form a Gaussian curve, the very high temperature areas will cold down faster than the low temperature areas. After a few seconds the old laser impact point can still be detected but past 10 seconds, in the best case, the hot area was too diffuse to see the dead center of the hot area. Note that the hot area will decrease to a temperature nearly equal to the object initial surface temperature plus 1 or 2 degrees and after this cold down at a very slow speed. It takes more than 5 minutes to some plastic samples to return to the surface temperature they used to have before the experiment.

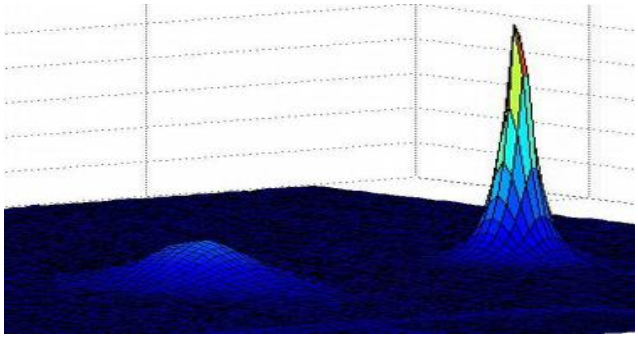


Figure 4: Left : old laser spot (4 seconds). Right : new laser sport (0.1 second)

4. Conclusion

For now we are hopping to finish the 3D calibration and dot-follow step to began large-scale experimentation with a actual moving system. The tests will began by a simple translation and we will add some complex movement pattern later if we have corrects results. We also need to try several types of objects made of many materials and of many forms to check if the point can still be followed.

This system is made to be able to build a 3D thermal map of the target object to give a human operator the possibility to see areas which are possibly compromised. By using thermal cameras we can use this system on a large among of objects, even some which are generally hard to study with normal camera like glass or any secular object.

References

[1] Alban Bajard, Numérisation 3D de surfaces métalliques spéculaires par imagerie infrarouge. phd, 20 novembre 2012

[2] Mohamed Belkacemi, Contrôle non destructif et numérisation 3D par thermographie active. phd, 6 december 2016

[3] I.Jorge Aldave, P.Venegas Bosom, L.Vega GonzálezaI. López de Santiago, B.Vollheim, L.Krausz, M.Georges. Review of thermal imaging systems in composite defect detection. Infrared Physics & Technology, Volume 61, November 2013, Pages 167-175

[4] Andrew W. Fitzgibbon. Robust registration of 2D and 3D point sets. Image and Vision Computing. Volume 21, Issues 13–14, 1 December 2003, Pages 1145-1153

[5] Usamentiaga, R.; Venegas, P.; Guerediaga, J.; Vega, L.; Molleda, J.; Bulnes, F.G. Infrared Thermography for Temperature Measurement and Non-Destructive Testing. *Sensors* 2014, *14*, 12305-12348.

High Dynamic Range Reflectance Transformation Imaging: an adaptative multi-light approach for the assessment of surfaces' visual quality

M. Nurit[a], G. Le Goïc[a], H. Favrelière[b], A. Mansouri[a]
[a]ImViA Laboratory, Université de Bourgogne Franche-comté, Dijon, France,
[b] Symme Laboratory (EA4144), Université Savoie Mont-Blanc (USMB), Annecy, France

ABSTRACT

Visual inspection of surfaces is a complex sensorial process that depend on many factors related to the observer[1-3] (as tiredness, past experience, expectations) and the light environment. In order to reduce the intrinsic variability associated with this kind of approach and improve the robustness of the analysis of the appearance quality of manufactured surface, we propose to couple the implementation of an innovative multi-imaging method called Reflectance Transformation Imaging (RTI) with the High Dynamic Range (HDR) processing. This coupling allows to avoid the low dynamic of the cameras and consequently insures a faithful estimation, in each pixel of the local angular reflectance of the inspected surface, especially very shiny ones. Beyond the visualization, further analysis such as as saliency maps are to confirm the pertinence of the proposed methodology.

Keywords: HDR Imaging, RTI imaging, Visual saliency, Visual appearance, Quality inspection

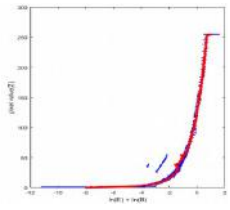
1. RTI FOR SURFACE QUALITY INSPECTION

RTI is a multi-lighting imaging method that provides access to multidimensional information of the inspected surface (Image relighting, slopes, curvatures and 3D mappings, visual saliency cartographies, etc.). These data are very valuable for evaluating and describing the visual quality of a surface, especially in the case of manufactured products [4]. The principle of acquisition is to vary the direction of the lighting source, the camera being positioned orthogonally to the inspected surface. Thus, in each point (pixel), an evaluation is obtained through a set of discrete values corresponding to the different illumination directions of the local angular reflectance.

A limitation of this approach is related to the range of measurement of the light response of the inspected surface. Indeed, conventionally, an image (which is dynamically limited by the sensor itself) is taken for each direction of illumination, called LDR image (Low Dynamic Range). However, in case of non-Lambertian surfaces exhibiting complex and/or heterogeneous local behavior, the LDR information is largely insufficient to describe fittingly the real physical response of each point of the inspected surface. This dynamic limitation also makes it difficult to correlate RTI data with human perception results[5-6]. It's therefore a major problem for the assessment of the visual quality of observed surfaces, and more globally for industrial inspection.

We propose to address this limitation by an auto-adaptive implementation of the HDR processing for RTI data-images. The principle is to automatically adapt the number of the acquired images as well as the exposure time values allowing the reconstruction of the HDR image for each of the acquisition directions, i.e for each of the illumination directions. This methodology is presented in section 2.

2. COUPLING HDR IMAGING WITH RTI: PROPOSITION OF AN ADAPTIVE METHODOLOGY

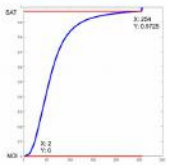


response Function (CRF)

Step 1. We determine the pixels that we will observe (100 to 5000 pixels distributed on the image), which is a good trade-off between computation time consumption (large number of pixels) and noise domination (small number of pixels) . Several images are captured to obtain a variation of radiance and some pixels must be contained in the working range of the camera. Then we have to find the solution of the following equation[7]:

$$0 = \sum_{i=1}^{N}\sum_{j=1}^{P} w(Z_{ij})[g(Z_{ij}) - lnE_i - lnB_j]^2 + \lambda \sum_{z=Z_{min}+1}^{Z_{max}+1} [w(z)g''(z)]^2$$

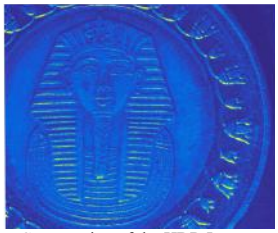
g(Zij) is the camera response function (CRF) to be determined. *Zij* the value of i^{th} pixel with the j^{th} exposure time. *Bj* is the j^{th} exposure time and *Ei* the unknown irradiance of the i^{th} pixel. w a weighting function.



Adaptive exposure time estimation and LDRs image capture

Step 2. At each angular position acquisition, exposure times for LDR acquisitions are adaptively determined. 3 parameters are used (fixed): The level of noise and the level of saturation of CRF, and a pixel threshold (percentage of saturated pixels that is considered acceptable). A first image is captured with an arbitrary exposure time (user choice). Then, and in a recursively manner as long as the pixel ratio properly exposed (not saturated) is too low, another image is captured with a longer and/or shorter exposure time determined automatically by using the CRF[8] (equation below):

$$\Delta_{shorter} = \frac{CRF^{-1}(NOI)}{CRF^{-1}(SAT)} * \Delta t \qquad \Delta_{longer} = \frac{CRF^{-1}(SAT)}{CRF^{-1}(NOI)} * \Delta t$$


Construction of the HDR Image

Step 3. We have, for each angular position i, a vector of images I with the j[th] element acquired with the j[th] exposure time in the vector B. With our set of images, we should apply a predicate to compute a weight matrix M. HDR images can then be calculated for each angular position.

$$M_i(x,y) = \sum_{j=1}^{|B_i|} \begin{cases} \text{if} & 0 < I_{i,j}(x,y) < 255, \quad 1 \\ \text{else} & \quad 0 \end{cases}$$

$$H_i(x,y) = \sum_{j=1}^{|B_i|} \frac{I_{i,j}(x,y) * B_{i,1}}{B_{i,j}} * \begin{cases} \text{if} & M_i(x,y) > 0, \quad \dfrac{1}{M_i(x,y)} \\ \text{else} & \quad \dfrac{1}{|B_i|} \end{cases}$$

Table 1: Coupling RTI with HDR Imaging: main steps of the method

3. RESULTS AND CONCLUSION

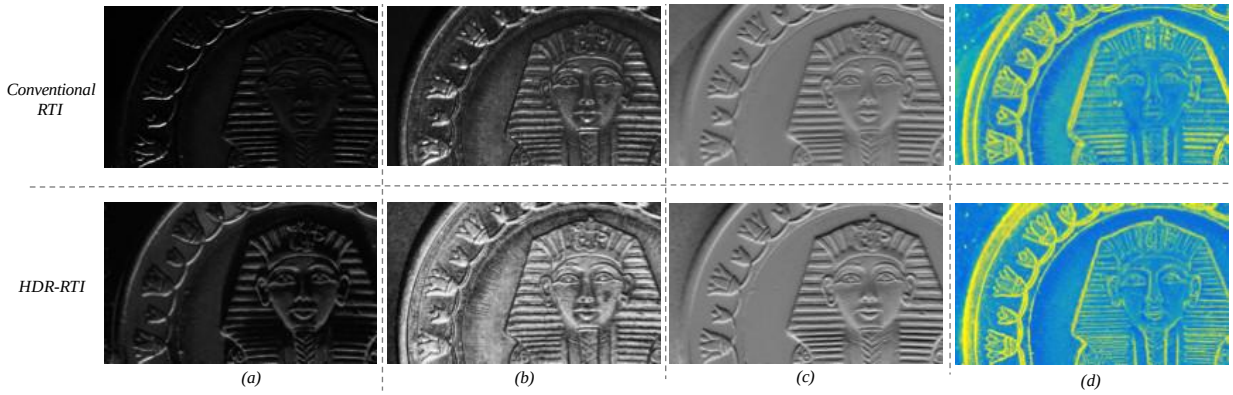


Figure 1. RTI results: Comparison between the conventional RTI and HDR-RTI acquisitions (from tone-mapped data).
(a) Image reconstruction for a raking incident light *(b)* Image reconstruction for a 65 degrees incident light *(c)* slope mapping *(d)* Visual saliency map

The comparison between results obtained from conventional RTI and HDR-RTI shows significant differences. HDR-RTI provides a complete description of the local reflectance in terms of dynamic. Thus, and as expected, the coupling of RTI with HDR imaging allows a better estimation of the local reflectance for raking and normal lighting, as it decreases strongly the number of saturated (low or high) information by adapting the exposition during the acquisition process. Moreover, other RTI modalities, derived from the local reflectance estimation, will also be affected. However, it can be noticed in figure 1 (slopes and saliency maps) that the changes are less visible than expected. This is because it is necessary to adapt the calculation methods to the new type of acquisition data (HDR).. It is a perspective of this work, in progress, which will be developed in the full version of the article.

REFERENCES

[1] Baudet, N., Pillet, M., and Maire, J., \Visual inspection of products: a comparison of the methods used to evaluate surface anomalies," International Journal of Metrology and Quality Engineering 2(1), 31{38 (2011).
[2] Guerra, A. S., Metrologie sensorielle dans le cadre du contr^ole qualite visuel, PhD thesis, Universite de Savoie - Laboratoire Symme (2008).
[3] Debrosse, T., Pillet, M., Maire, J., and Baudet, N., \Sensory perception of surfaces quality," Proceedings of KEER 2010 - International conference on Kansei engineering and emotion research , Paris, France , 1-11 (2010).
[4] Le Goic, G. and Samper, S., \Systeme de detection d'anomalies d'aspect par la technique PTM," (2011).
[5] Nayar, S. K. and Mitsunaga, T., \High dynamic range imaging: Spatially varying pixel exposures," in [CVPR], (2000).
[6] Debevec, P. E., Reinhard, E., Ward, G., and Pattanaik, S. N., \High dynamic range imaging," in [SIGGRAPH Course Notes], (2001).
[7] Debevec, P. E. and Malik, J., \Recovering high dynamic range radiance maps from photographs," in [Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques], 369{378 (1997).
[8] Martnez, M. A., Valero, E. M., and Hernandez-Andres, J., Adaptive exposure estimation for high dynamic range imaging applied to natural scenes and daylight skies," Appl. Opt. 54, B241{B250 (Feb 2015).

MICRO-EXPRESSIONS RECOGNITION PRESENTATION

R. Belaiche, C. Migniot, D. Ginhac, F. Yang

ImViA EA 7535, Univ. Bourgogne Franche-Comté, Dijon, France

ABSTRACT

Machine learning has known a tremendous growth within the last years, and lately, thanks to that, some computer vision algorithms started to access what is difficult or even impossible to perceive by the human eye. Our objective is to create an artificial intelligence able to automatically detect and classify facial Micro-Expressions(ME) of emotions. Being able to recognise ME offers many applications such as trustworthy lie and deception detectors. ME would also be useful for repressed felling detection, as ME can happen when the person herself isn't conscious of her feelings.

1. INTRODUCTION

Facial expressions offer important benchmarks in every day's social interactions. Most people are familiar with macro facial expressions, however, few people are aware of the existence of micro-facial expressions [1][2], and even fewer know how to detect and recognize said micro-expressions. Initially discovered by Haggard and Isaacs [3], micro-expressions are a type of involuntary facial expressions that are extremely fast and of very low intensity. Their duration is within 1/4 seconds, which makes their localization and analysis rather complicated tasks. Micro-expressions(ME) can occur in two situations: conscious suppression and unconscious repression. Conscious suppression happens when a person intentionally tries to stop themselves from showing their true emotions or try to hide them. Unconscious repression occurs when the subject himself does not realize their true emotions. In both cases, micro-expressions betray the subject's real emotions independently from his awareness of their existence. In this paper, we re-visit the baseline results for Micro and Macro-Facial Expressions(M/M-FEs) classification for the $CAS(ME)^2$ dataset [4] using the *Local Binary Pattern Three Orthogonal Planes* (LBP_TOP) operator as feature extractor.

The paper is organized as follows : we describe the feature extraction method in section 2. Experiments are presented and discussed in Section 3 and conclusion is given in Section 4.

2. FEATURE EXTRACTION METHOD

In order to recognize Micro Expressions, we have to go through two steps: feature extraction and classification. M/M-FEs can be described with the help of a spatio-temporal local texture descriptor used to get the features giving information on the pixel content of the whole images as they change over the time.

We use the LBP_TOP operator, which is the baseline descriptor used as reference in most papers studying micro-expressions [5][4] to describe M/M-FEs videos. The LBP operator is a type of visual descriptor that was originally designed for texture description [6]. The general idea is to threshold a small area around each pixel in order to build a binary code. This code is obtained by comparing neighbour pixel values with the center pixel: values superior or equal to the center pixel's value get assigned a 1 while smaller values get assigned a 0. The choice of the surrounding area directly affects the kind of edges it is possible to detect in an image. For pixel neighborhoods referring to *N* sampling points on a circle of radius *R*, we generally use the notation $LBP_{N,R}$, whose value for a pixel c can be given by :

$$LBP_{N,R} = \sum_{p=0}^{N-1} t(g_p - g_c)2^p \quad (1)$$

Here g_c represents the gray value of the center pixel c while g_p represents the gray values of equally spaced pixels on a circle of radius R, t defines a thresholding function $t(x) = 1$ if $x \geq 0$ and $t(x) = 0$ otherwise. The feature vector representing an input image is calculated by extracting the histogram distribution of the LBP.

We can consider LBP as texture primitives that include different types of curved edges, spots, flat areas, and so on. For an efficient facial representation, images usually get divided into local blocks from which we extract the LBP histograms and concatenate them into an enhanced feature histogram [7]. Local texture can then be described using said histograms of the binary values for a block in the image. The number of blocks and the size of each block determines the level of retained spatial information.

The conventional LBP only serves for spatial data in 2D images. To describe data from the 3D spatio-temporal domain, the basic LBP is extracted from the three planes XY,

XT and YT for each pixel as shown in Fig.1. The resulting three histograms are then concatenated into a feature vector describing the video. After being originaly proposed for dynamic textures description [8], it was first used for micro-expressions recognition by Pfister *et al.* [5] .

The feature vectors extracted from XT, YT and XT are then concatenated and given as input to an SVM with an RBF kernel for classification.

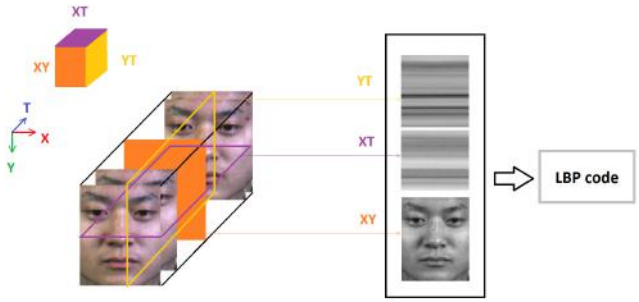


Fig. 1. Illustration of a spatiotemporal volume of a video [4]

3. EXPERIMENTS AND DISCUSSION

3.1. Dataset presentation

The number of scientific papers dealing with the automatic analysis of micro-expressions is rather limited. One of the reasons for it can be attributed to the lack of datasets containing real micro-expressions. Fortunately, this is beginning to change and new foundations are being laid, with new datasets relating spontaneous micro-expressions. As a matter of fact, the dataset we work on,$CAS(ME)^2$ dataset [4], is one of the few datasets to present annotated videos of spontaneaous M/M-FEs of different subjects. 22 participants in total were filmed while watching different kinds of excitation videos. Each subject was informed that their monetary rewards would be reduced if they produced too many noticeable facial expression.

The dataset was originally proposed for automatic M/M-FEs spotting and recognition, and while micro-expressions spotting has been getting good results [9], their recognition still offers many challenges.

$CAS(ME)^2$ offers 2 kinds of annotations for MEs, the first one is done according to the facial muscle movements based on the *Action Units* (AU) following the Facial Action Coding System (FACS) proposed by Ekman. The second annotations are based on self-reported emotions from the candidates. These two annotations are not in agreement for all the videos. Furthermore, in some cases, emotional states and the annotations based on AU are contradictory (some people would show a negative facial expression in front of a happiness-inducing video). 24.05% of the facial expressions are classified as *others* (*i.e.* where related AU are not discriminative). The low intensity of the ME is one of the biggest challenges for ME classification.

M/M-FEs

true \ pred	*pos*	*neg*	*surp*	*other*
pos	63.1	12.8	4.3	13.4
neg	17.1	45.6	73.9	41.5
surp	0	0	0	0
other	19.8	41.6	21.7	45.1
Acc.	**48.09**			

Table 1. Emotional M/M-FEs classification confusion matrix and accuracy. Results are expressed in percentage (%)

3.2. Model validation protocol

Classification was tested following the *Leave One Subject Out* (LOSO) cross-validation protocol: one subject's data is used as a test set in each fold of the cross-validation. This is done to better reproduce actual use conditions where the encountered subjects are alien to the model when it was trained. Older studies would use k-fold cross-validation; however, this would result in a severe case of overfitting as the accuracies on the training sets would be much higher than with LOSO. This can be attributed to the fact that samples from the same subject would be present in both the training and testing sets. Considering the fact that the same subject can show the same expression many times (which may cause that expression to belong to the training and the test sets at the same time), and that some subjects can be more inclined to show a specific type of emotion more often, using the LOSO protocol seems to be the most rigorous option.

3.3. Results

We divided the images into 5 by 7 blocks before applying LBP_TOP on each 3-Dimensional block. The final results of M/M-FEs are given by Table 1 . Accuracy rates describe how reliable the classifiers are at predicting emotional states correctly (presence rate of true positives) following the LOSO cross-validation protocol.

We can see that the model is not competent for recognising surprise. The reason for that is that only 06.74% of the dataset shows suprised M/M-FEs while 32.55% of the dataset amounts to positive, 36.66 is negative and 24.05% is accounted as other.

4. CONCLUSION

MEs are extremely useful for punctual emotion classification. From simple videos it is possible to extract analytical data

describing a Micro-Facial Expression. Future works may include the use of other datasets along the $CAS(ME)^2$ one. We can also imagine a fusion of MEs and other modalities to describe emotional states.

5. REFERENCES

[1] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, 1969.

[2] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*, WW Norton & Company, 2009.

[3] E.A. Haggard and K.S. Isaacs, *Methods of Research in Psychotherapy*, Springer, 1966.

[4] F. Qu, S.J. Wang, and W.J. Yan et al., "Cas(me)^2: A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Trans. on Affective Computing*, 17 January 2017.

[5] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Recognising spontaneous facial micro-expressions," *IEEE Inter. Conf. on Computer Vision*, 2011.

[6] T. Ojala, M. Pietikinenand, and T. Menp, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2002.

[7] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2006.

[8] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2007.

[9] Y. Han, B. Li, Y. K. Lai, and Y. J. Liu, "Cfd: A collaborative feature difference method for spontaneous micro-expression spotting," *25th IEEE ICIP*, 2018.

Doctoral day SPIM Vol: 4 (2019)

To an efficient method to estimate trifocal tensor based on lines
Doctoral day 2019

Daniel Braun*, Pascal Vasseur* and Cédric Demonceaux†
* Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76 000 Rouen, France
† ImViA ERL VIBOT CNRS 6000, Université de Bourgogne-Franche-Comté, 71 200 Le Creusot, France

***Abstract*—This paper presents a new efficient approach for the trifocal tensor estimation in three view geometry based on lines, to propose a viable alternative to the points based method. Current state of the art approach using lines requires to simultaneously have 13 lines visible from the three views, when the points method only requires 7 points. Our objective here is to reduce to 8 the number of required lines by imposing two angles of the camera rotation matrix as known. The validation of the method is done on simulated data in order to estimate its robustness to noise and to uncertainties of the camera angles measurement. Even if this approach is specific to a particular case where we have access to camera angles, it remain an interesting solution for all mobile systems that are generally equipped with Inertial Measurement Unit (IMU).**

***Index Terms*—Trifocal Tensor, Multiple View Geometry, Pose Estimation**

I. Introduction

In computer vision, the 3D reconstruction and pose estimation are head research subject for robotic application to localize a robot with embedded intrinsic sensors. Yet most of the approaches are limiting themselves to two-view geometry with the fundamental matrix that bind all projective geometric relations between the two cameras composing the system. Still it can be extended to three view-geometry by introducing a new object called the trifocal tensor that serves an analogous role as the fundamental matrix among the three views. Its main advantage is its capability to transfer points or lines from two views to a third, which is highly beneficial to perform correspondences over multiple views.

Hartley and Zisserman [1] have shown that the trifocal tensor can be estimated by matching triplets of points, lines or any combination of points and lines. Nonetheless, the most common methods are used with sets of points [2] since it is easier to implement than line based methods. Yet, lines are stronger scene descriptors than points, so our objective is then to propose an efficient approach using lines to equal or overcome the one with points.

In this study, we will introduce the trifocal tensor, how it can be defined and what are its mains properties. Then we will present our algorithm, which is the constraint made to simplify the system and how it impacts the trifocal tensor. We will finish by presenting our experimental approach and conclude on the obtained results. Since it is still a work in progress, some more experimentations will have to be done to decide on the method viability.

II. Trifocal tensor

Throughout the paper, we will use the following notations : vectors are represented by lowecase bold ($\boldsymbol{v}$), matrices by uppercase (M) and scalar by lowercase (s). The left cross product is denoted with the 3×3 skew-symetric matrix $[\boldsymbol{a}]_\times$ such that $\boldsymbol{a} \times \boldsymbol{b} = [\boldsymbol{a}]_\times \boldsymbol{b}$.

A. Definition

The trifocal tensor is generally represented by a set of three 3×3 matrices $\{T_1, T_2, T_3\}$. It is gathering all geometric relations binding three views that are independent to scene structure. The system is defined in its canonical form by three projective cameras $P_1 = [I \mid 0]$, $P_2 = [A \mid \boldsymbol{a}_4]$ and $P_3 = [B \mid \boldsymbol{b}_4]$ such that

$$T_i = \boldsymbol{a}_i \boldsymbol{b}_4^T - \boldsymbol{a}_4 \boldsymbol{b}_i^T \text{ for } i = 1, .., 3 \tag{1}$$

where $\boldsymbol{a}_4$ and $\boldsymbol{b}_4$ are the epipolar lines of the second and third camera respectively from the first camera and the vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are the i^{th} columns of the rotations matrix A and B respectively.

The trifocal tensor is composed of 27 elements, which are 26 independent ratios up to a common scale. Yet it only has 18 degrees of freedom (dof), considering that each camera has 6 dof in the projective world frame. It means that the system can be completely described by 18 parameters.

B. Trilinearity constraint

According to Hartley and Zisserman [1], if a line is visible in the three cameras views, it exists a relation linking the three projected lines called the trilinearity constraint. It is defined by the following relation :

$$\boldsymbol{l}_i^1 = (\boldsymbol{l}^2)^T T_i \boldsymbol{l}^3 \text{ for } i = 1, 2, 3 \tag{2}$$

with $\boldsymbol{l}_i^1$ representing the i^{it} coordinates of the line $\boldsymbol{l}^1$ in the first camera. $\boldsymbol{l}^2$ and $\boldsymbol{l}^3$ are the coordinates of the same line respectively in the second and the third camera and T_i is the i^{th} matrix of the trifocal tensor.

Similar relations exist between points and any lines and points combination [1] but it is not the subject of this paper.

III. Algorithm

A. Rotation constraint

In order to reduce our system complexity, we impose a constraint on the camera rotations. In that way, we assume we have access the camera's attitude, i.e. the pitch and roll rotations, leaving the camera only depending on the yaw rotation. The new constraint system can be expressed has :

$$P_1 = \left[I \mid 0\right],\ P_i = \left[R_i^z \mid -R_i^z \boldsymbol{t}_i\right] \text{ for } i = 2, 3 \qquad (3)$$

where R_i^z is the rotation matrix around the z axis of the i^{th} projective matrix and $\boldsymbol{t}_i$ is the i^{th} camera position in world coordinates.

Based on equation 1 and the new constraint system, the number of elements of the trifocal tensor has been reduced from 27 to 21 with only 16 linearly independant elements, plus the overall scale.

B. Trifocal tensor estimation

The equation 2 can be rewritten as :

$$\left((\boldsymbol{l}^2)^T \left[T_1, T_2, T_3\right] \boldsymbol{l}^3\right) [\boldsymbol{l}^1]_\times = \mathbf{0}^T \qquad (4)$$

From this relation, we can extract two independent equations, which means that 8 triplets of lines a necessary to get the 16 equations and fully solve the system, up to scale.

The equations system can be written as $A\boldsymbol{t} = \mathbf{0}$ with A a 16×17 matrix and $\boldsymbol{t}$ a 21 vector containing all the entries of the trifocal tensor. The scaling factor is imposed with the constraint $\|\boldsymbol{t}\| = 1$ and This system is solved using the Singular Value Decomposition (SVD) algorithm that provides the best estimation for $\boldsymbol{t}$.

C. Estimate the camera pose

Once the trifocal tensor has been estimated, the camera pose can be easily retrieved by using the relation 1, providing us an overdetermined system.

IV. Experiments

The validation of the method is performed on synthetic data composed of a set of lines automatically generated and defined by there two end points centered around the world's origin.

A. Robustness to noise on image points

To evaluate our method robustness to noise, we apply white gaussian noise on the image points with a standard deviation σ varying from 0 to 3 pixels. For each level of noise, the error is average over 100 executions.

Since we are working on line, it is interesting discuss on how to properly apply the noise. It would have been possible to simply apply noise to the two end points but it could result in a very disadvantagous configuration if the noise moves the two points in opposite direction. To have closer results to what can perform line detection methods [3], [4], each line is considered to be composed of several points depending of its length. The noise is applied on the points and the new line is computed by linear regression.

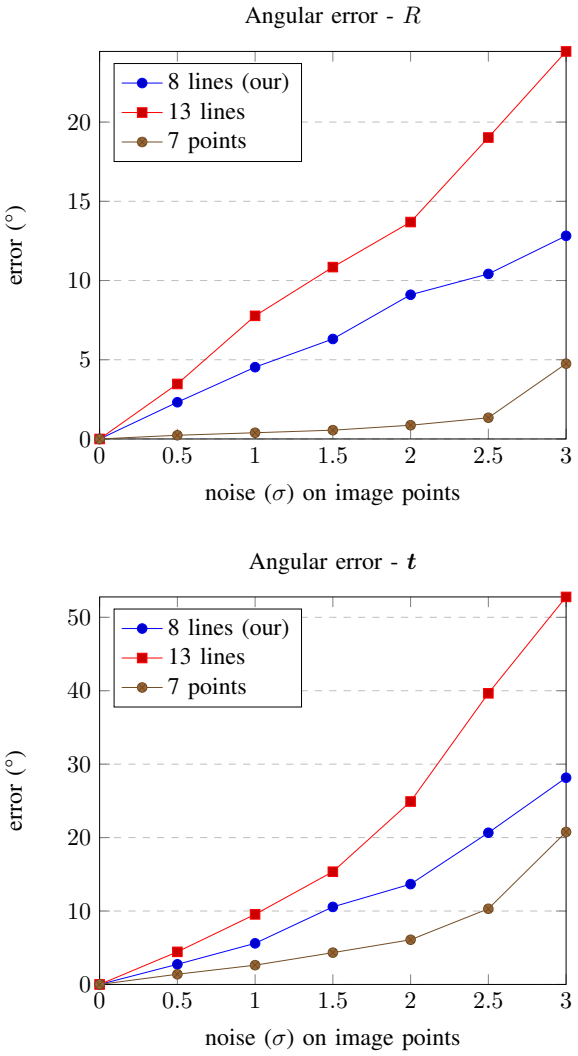


Fig. 1. Average angular error for the camera rotation (R) on the top and for the translation direction ($\boldsymbol{t}$) on the bottom, depending on the gaussian noise added to the image points. Comparison of our 8 lines method to the 13 lines and the 7 points methods.

In figure 1, we compare the behaviour of the 8 lines, 13 lines and 7 points algorithms to image points noise. Even if the 7 points method stays the more stable, our approach with 8 lines is outperforming the 13 lines algorithm.

B. Robustness to noise on angle from IMU

In a second time, we study the impact of the noise on the input angles acquired by the IMU. As before, a white gaussian noise ($\Delta\theta$) is applied to the two rotation angles in degree. Its performances are compared in Figure 2 to the 13 lines algorithm and to the 8 lines without noise on the IMU. Without surprise, the noise on the IMU is impacting the pose estimation accuracy. Yet, for a noise on the image points of $\sigma > 1.5$ and an angular error of the IMU of $\Delta\theta = 0.25$, we have better results than the 13 lines algorithm.

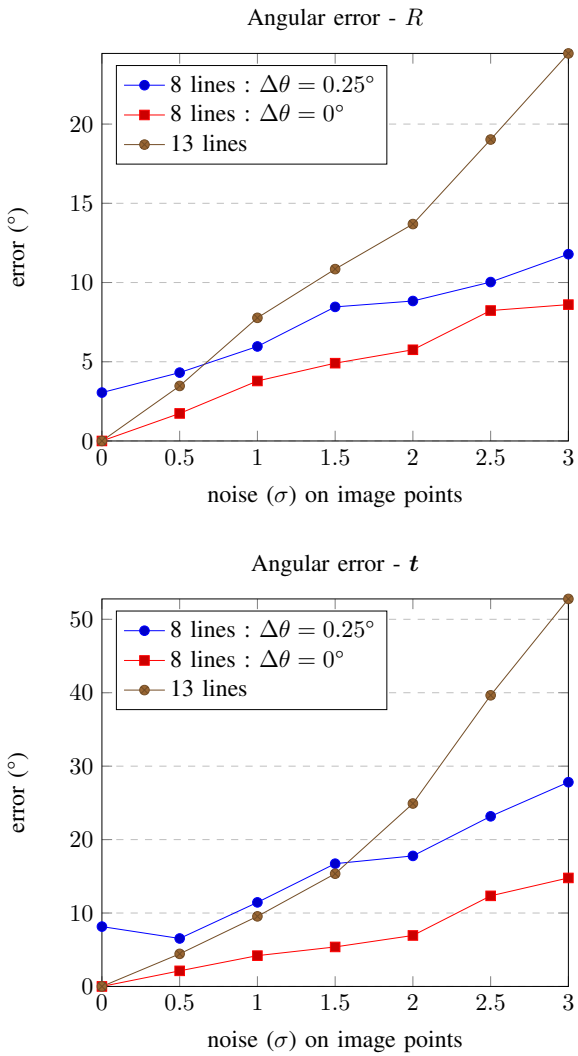


Fig. 2. Average angular error for the camera rotation (R) on the top and for the translation direction ($\boldsymbol{t}$) on the bottom, depending on the gaussian noise added to the image points. Comparison of the 8 lines method with noise on the IMU angles ($\Delta\theta$) to the 8 lines without noise on the IMU and to the 13 lines method.

V. Conclusion and Future Work

Considering the first experimental results, the estimation of the trifocal tensor with 8 lines and two given camera angles is still sensitive to noise and not as efficient as the points method. Yet, it outperformes the 13 lines algorithm, even with noise on the IMU measurement and on the image points combined.

In future work, we will evaluate our approach performances on a real world benchmark. Its performances will be compared to the 7 points and 13 lines algorithm.

References

[1] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second ed., 2004.

[2] L. F. Julià and P. Monasse, "A critical review of the trifocal tensor estimation," in *Image and Video Technology* (M. Paul, C. Hitoshi, and Q. Huang, eds.), (Cham), pp. 337–349, Springer International Publishing, 2018.

[3] Z. Wang, F. Wu, and Z. Hu, "Msld: A robust descriptor for line matching," *Pattern Recognition*, vol. 42, no. 5, pp. 941–953, 2009.

[4] L. Zhang and R. Koch, "An efficient and robust line segment matching approach based on lbd descriptor and pairwise geometric consistency," *Journal of Visual Communication and Image Representation*, vol. 24, no. 7, pp. 794–805, 2013.

Acknowledgment

To conclude this proceeding we would like to thank all the participants who took the time to compose their papers as well as providing structured and consistent reviews.

Thank to the students :
Abir Zanzouri, Ahmad Zawawi Jamaluddin, David Strubel, Daniel Braun, Marc Blanchon, Thibault Clamens, Thomas Herrmann and Yifei Zhang
who took on their time to organize this event.

Thanks to SPIM for the financial support

We would like to address a special thank to Nathalie Choffay for her help.

We couldn't end this proceeding without acknowledging David Fofi, Cédric Demonceaux and Franck Marzani for their supports.