Detecting Swimming Pools in 15-Minute Load Data

Sebastian Burkhart, Andreas Unterweger, Günther Eibl and Dominik Engel
Center for Secure Energy Informatics, Salzburg University of Applied Sciences, Puch bei Hallein, Austria
Email: firstname.lastname@en-trust.at

***Abstract*—The ability to detect appliances in load data highly depends on the resolution of the data. While a lot of related work exists on detecting appliances in second or sub-second granularity load data, in this paper, we detect swimming pools through their filter pumps in load data with the 15-minute granularity prescribed by the European Union for smart meters. We model the filter pump based on exemplary measurements and describe a prototypical algorithm to extract the filter pump's consumption from the aggregated mains signal of a real-world household. We evaluate pool detection performance with different classifiers on a data set with 843 households, where the information on the existence of a swimming pool is available. We achieve 94.8% detection accuracy with a precision of 68.5% with an off-the-shelf classifier. Decreasing the temporal resolution in several steps to 8 hours negatively affects the recall while the precision stays at the same level. We find that these results raise privacy concerns even at the minimum temporal resolution of smart meter data that is legally required in the European Union.**

I. Introduction

Non-intrusive load monitoring (NILM) aims at disaggregating load data [1], i.e., splitting consumption patterns based on the devices which incurred them. NILM algorithms are typically applied to data with relatively high temporal resolution [2]–[4]. High-resolution NILM approaches typically use sampling frequencies in the order of kHz, while so-called low-frequency approaches typically measure at frequencies in the order of 1 Hz [5], [6]. Even lower resolutions should also be investigated since the European Union requires 15 minutes as the minimum measurement interval of smart meters [7]. While it is known that frequencies of the order of 1 Hz pose privacy issues [8], [9], this also needs to be investigated for lower frequencies.

Low-resolution approaches can be divided into unsupervised and supervised approaches. Unsupervised approaches typically use some variant of hidden Markov models like Conditional Factorial Hidden Semi-Markov models [10] or difference Factorial Hidden Markov Models [11]. Supervised approaches for the most part still use a variant of the method that has originally been outlined in [12].

It is well-known that the performance of low-resolution methods decreases when multiple events of different appliances occur within the measurement interval. This situation occurs when (i) many appliances are used in a household [10]; and, of course, when (ii) the measurement interval increases. The decrease of resolution has been shown to lead to worse performance (and higher privacy) for methods relying on edge detection [13].

Compared to the wealth of research on methods using low or high resolution, research about very coarse resolutions, i.e., in the order of 15 minutes, is rare. While occupancy detection methods largely use low resolutions, the approach in [14] uses 1-minute resolution data. Similarly, in a recent paper, data with a granularity of 15-minutes has been analyzed for holiday detection [15]. In [16], load data with resolutions of one minute, 15 minutes and one hour are used to disaggregate air conditioning appliances to detect households with air leaks. This approach works well for one-minute data but achieves a poor performance for the lower resolutions.

In this paper, a new algorithm is developed that for the first time enables to detect an appliance (a swimming pool) in low-resolution data ($\Delta t \geq 15$min) of regular households. This would not be possible by using low-resolution NILM approaches due to the comparably high number of appliances present in normal households and the coarse time interval.

It must in turn be noted that the practical use of this method for applications such as home automation is limited. The proposed method requires a rather long measurement history and it only states whether or not the appliance is present. While the accuracy is not perfect, the results clearly can be exploited, e.g., by marketing applications. Most importantly, this analysis shows that even the coarse 15-minute time intervals can possibly be misused, resulting in a privacy decrease for the end-user.

This paper is structured as follows: In Section II, we show which effect swimming pools have on load profiles and how to detect this effect. We apply the developed detection algorithm to a large set of real-world load profiles in Section III and examine the influence of lower temporal resolutions on the pool detection performance in Section IV. We conclude the paper in Section V.

II. Swimming Pool Model

In this section, we describe how to detect swimming pools in consumption data with 15-minute granularity. To motivate the features used in our detection algorithm, we first describe which electrical appliances are typically required to operate a swimming pool and how they behave in terms of power consumption over time.

A. Filter pump sample data

Usually, a swimming pool requires a filter pump in order to clean the water and prevent accumulations of algae. Since filter pumps typically run at regular time intervals over several months consuming several hundred Watts of electrical power, it is plausible to expect easily detectable patterns in the load data.

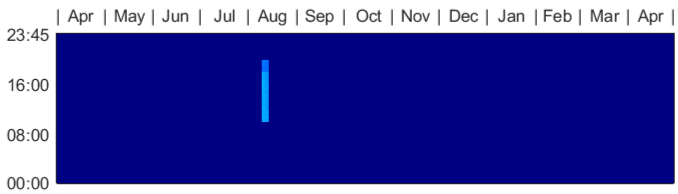
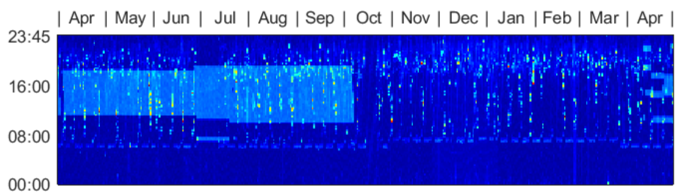
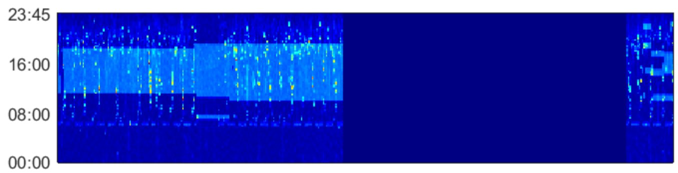


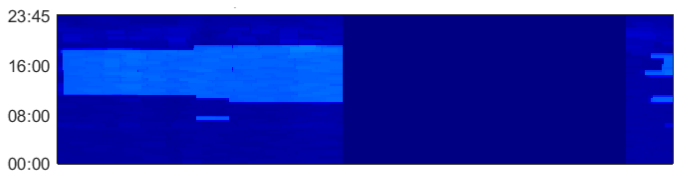
Fig. 1. Load profile of a filter pump over five days.



(a) Image construction from the normalized load profile with a color palette

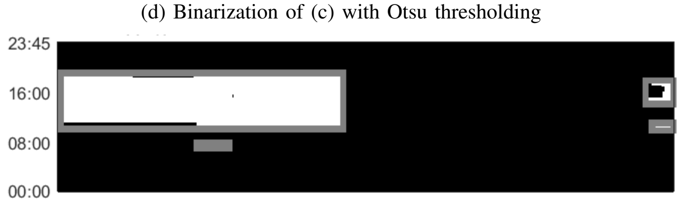


(b) Removal of the fall and winter months from (a)



(c) Opening of (b) with a rectangular $10 \cdot 2$ structuring element

(d) Binarization of (c) with Otsu thresholding



(e) Detection of rectangles as bounding-boxes of 8-connected areas in (d)

Fig. 2. Pre-processing steps: A load profile visualization (a) is pre-processed ((b)-(d)) to simplify rectangle detection (e). Rectangular patterns in the spring and summer are likely to indicate additional consumption due to a filter pump.

In order to test these assumptions, we installed a measurement device (Allnet ALL3075v3) to monitor the consumption of an existing filter pump (Miganeo Speedclean 8500) and a saltwater system (Intex ECO8220) for a period of five days in the month of August. Fig. 1 visualizes the measured power consumption with 15-minute granularity. The X axis shows the months from April to May of the following year (which is identical to the measurement period of test data as explained in Section III). The Y axis shows the time of day in 15-minute intervals. Dark tones indicate low or zero power, bright tones indicate higher power, with the maximum being 480 Watts.

From the measurements visualized in Fig. 1, we observe the following properties:

- **Significant power consumption**: The pump consumes either near-zero or about 450 Watts with relatively little variation;
- **Regular time interval**: Due to an automatically scheduled on-off mechanism, the pump has identical operating times each day.

In addition, we assume that the filter pump is only operational during the warmer period of the year, i.e., the consumption pattern of the filter pump is expected to last as long as this period. Furhermore, we assume that these three properties apply to typical setups with filter pumps and therefore allow detecting swimming pools with reasonable accuracy. This leads to the assumption that households without swimming pools lack filter-pump-like consumption patterns.

B. Pre-processing

Based on the three properties discussed in the previous section, we propose a simple algorithm to extract features from load data suitable to distinugish filter pumps from other appliances. Our algorithm consists of five steps which are illustrated in Fig. 2:

- **Normalization**: 15-minute load data is normalized by dividing by the maximum value. Fig. 2a depicts a visualization of the data with *Matlab*'s *jet* color map. The same visualization resulted in Fig. 1 for our filter pump sample data;
- **Scoping**: Based on the assumptions from Section II-A, filter pumps are only active during warmer months, which, in the case of Austria, are the spring and summer months. Thus, the months between October and March are removed as depicted in Fig. 2b;
- **Opening**: As discussed in Section II-A, filter pumps are active during the same time of day and over long periods of time, resulting in rectangular-like shapes as is apparent (for humans) in Fig. 2b. To detect and isolate these rectangular structures, we need to do both, preserve the structure itself and remove consumption characteristics from other appliances. In image processing, this can be achieved by a morphological opening [17]. We use a rectangular $10 \cdot 2$ structuring element, representing a ten-day load of half an hour duration, similar to the structure and properties shown in Fig. 1. The opening operation yields Fig. 2c;
- **Binarization**: In order to separate the rectangular shapes (supposedly representing the consumption of a filter pump) in the foreground from the background, we binarize the pre-processed data using Otsu's thresholding algorithm [18]. Through that, the foreground (white) is distinct from the background (black), as illustrated in Fig. 2d.
- **Rectangle detection**: The position and size of the individual rectangular shapes are not yet accessible. In order

to obtain this information, we compute bounding boxes (illustrated as gray borders in Fig. 2e) around all foreground pixels which are connected to their neighboring pixels. This results in a list of rectangles together with their size and position.

Note that the detected rectangles by themselves are not the features used to classify whether or not a household has a swimming pool. In the next section, we will develop features from the properties of the rectangles to allow for such a classification.

III. Evaluation

In this section, we apply the pre-processing algorithm from Section II to a real-world data set. Before doing so, we describe the data set and our evaluation methodology.

A. Data set

The smart meter data used for evaluation has been collected between April 2010 and May 2011 from households located in the federal state of Upper Austria in Austria. The data has been collected in a study conducted by the *Energieinstitut* at the Johannes Kepler University Linz.

In addition to the electrical power consumption in 15-minute intervals collected by the smart meters, the data set contains demographic data from conducted surveys, including information on the existence of residential swimming pools per household. The sample contains 843 households, 64 of which have a pool.

B. Methodology

The algorithm proposed in Section II is designed to detect regions (bounding boxes) suspected to represent times where the filter pump of a swimming pool is powered on. To determine whether a filter pump is present, the original load data is first divided into $n+1$ parts – the n bounding boxes $B_i,\ i = 1, \ldots, n$ likely to represent filter pump activity from pre-processing; and the (one) remaining part of the load data, R. For example, Fig. 2e contains $n = 4$ bounding boxes (highlighted in gray) and the (black) remainder R outside of these bounding boxes.

Second, the following five properties are calculated from the load data L and the derived binarization L_{bin} from Fig. 2d so that they can be used as features:

- **Number of regions**: n;
- **Total region area**: Let w_i and h_i denote the width and height of the bounding box B_i, respectively. The total region area is calculated as $a := \sum_i w_i \cdot h_i$;
- **Average region coverage**: Let $f_i := \mathrm{cnt}(L_{bin}(B_i))$ denote the number of (white) foreground pixels within the bounding box B_i, where $\mathrm{cnt}(\cdot)$ computes the number of white pixels in the given area. The average coverage is calculated as $c := \frac{1}{a} \sum_i f_i$;
- **Median of region-medians**: $m_b := \mathrm{med}_i(\mathrm{med}(L(B_i)))$, where $L(B_i)$ denotes the load data in the region B_i and med is the median function;
- **Median of remainder**: $m_r := \mathrm{med}(L(R))$.

TABLE I

Performance of different classifiers for the 580 households in the *multi-region* case: The 5-nearest-neighbor classifier outperforms the uninformed ones in terms of precision.

Classifier	Accuracy	Precision
All-positive	10.5%	10.5%
All-negative	89.5%	–
SVM Gaussian	93.1%	66.7%
5-NN	94.0%	*68.5%*
1-NN	93.4%	66.7%

Third, all five properties are standardized so that they have a mean of 0 and a standard deviation of 1. This results in the five features $\tilde{n}$, $\tilde{a}$, $\tilde{c}$, $\tilde{m}_b$ and $\tilde{m}_r$.

Finally, these features are used to build a classifier to detect the presence of swimming pools distinguishing two cases: If the whole picture is covered by regions, i.e., $R = \{\}$, m_r is not defined. In this case, the corresponding household is considered not to have a swimming pool ($y = 0$). We refer to this case as *single-region*. In the opposite case, $R \neq \{\}$ or *multi-region*, we use one of several classifiers that are trained with the aforementioned five standardized features – $\tilde{n}$, $\tilde{a}$, $\tilde{c}$, $\tilde{m}_b$ and $\tilde{m}_r$. The performance of each classifier was assessed using the leave-one-out error because the low number of 64 households with pools (positive samples).

C. Results

In the data set, the *single-region* case applies to 263 households. 9 of these households have a swimming pool, whereas 254 do not have one. Since all these 263 households are classified as having no pools (negatives), the 9 households with a swimming pool are predicted as false negatives by the first step of our classification algorithm. Conversely, the remaining 254 true negatives are predicted correctly.

For the remaining 580 of the total 843 households, i.e., those for which the *multi-region* case applies, 525 have no swimming pool, whereas 55 have one. We trained the classifiers listed in Table I for these households as described in Section III-B. Since knowledge about a swimming pool is expected to correlate with properties such as large lots, high income and wealth, this information is considered privacy-sensitive due to the potential abuse. For example, as part of targeted marketing, advertisements may be sent only to households classified as having a swimming pool. Therefore, we use the precision as our main performance criterion.

It is clear from Table I that uninformed classifiers, i.e., those whose output (all positive or all negative) does not depend on the features, perform poorly in terms of precision. In contrast, the 5-nearest-neighbor classifier outperforms the all-positive classifier by about a factor of 6.5. Note that more sophisticated classifiers may perform even better. However, our goal was to show that low-resolution consumption information can be used to detect swimming pools in principle with reasonable accuracy.

The overall performance is determined by combining the *single-region* and *multi-region* classification results. When using the 5-nearest-neighbor classifier for the *multi-region* case,

		Predicted class		
		no pool	**pool**	**total**
True class	**no pool**	762 96.6%	17 31.5%	779
	pool	27 3.4%	37 68.5%	64
	total	789 100%	54 100%	843

Fig. 3. Confusion matrix for the 5-nearest-neighbor classifier: The overall performance of our pool detection algorithm, i.e., both, the *single-region* and the *multi-region* case, is 68.5% in terms of precision. The overall accuracy is 94.8%.

the overall performance for all 843 households is illustrated in Fig. 3 as a confusion matrix. Note that the overall precision does not differ from the precision in the *multi-region* case (see Table I), since *single-region* households are classified as having no swimming pool. However, the overall performance of an uninformed classifier is 7.6%, which is a factor of 9 less. In addition, the overall accuracy of 94.8% is slightly higher than that of an uninformed classifier, which is 91.8%.

IV. Lower Resolution

So far the considered time interval has been 15 minutes (original sampling rate). Now we study the influence of an increasing time interval (i.e., decreasing sampling rate) on the pool detection performance.

A. Experimental Setup

The method above was applied to data of lower temporal resolution which were obtained by averaging consecutive values in the original data. However, in order to ensure that every pixel of the resulting heatmap represents the same timespan, only certain time intervals were considered (see Fig. 4). The resulting heatmap contains the same number of days but is vertically squeezed. In section II-B a rectangular $10 \cdot 2$ structuring element (representing ten days and 30 minutes) has been used in the pre-processing step. Since the new time intervals are at least of twice the original length, the corresponding size of the structuring element only has a height of one pixel. This modified structuring element was used for all temporal resolutions but the original one.

For each time resolution previously chosen 5NN-classifier was evaluated using the leave-one-out error in the same way as for the original data.

B. Results

Analogously to the previous analysis the precision was chosen as the primary performance measure. Figure 4 shows

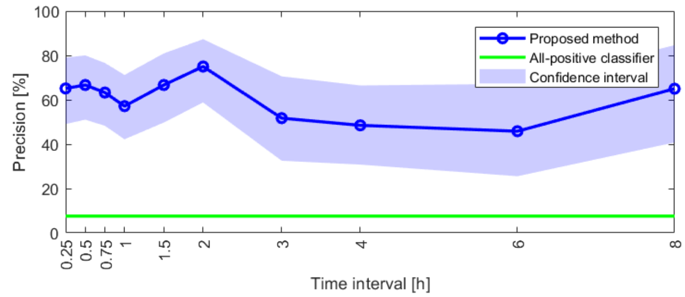


Fig. 4. The precision of the proposed method, though considerably higher than the all-positive classifier's precision, shows no statistically significant dependency on the time interval.

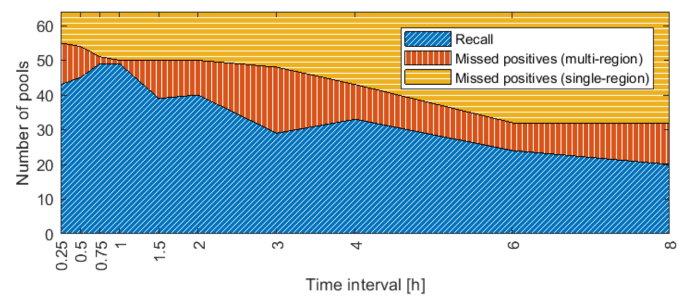


Fig. 5. Dependency of the recall on the time granularity. Two categories of missed swimming pools (positives) lead to a decrease with growing time interval.

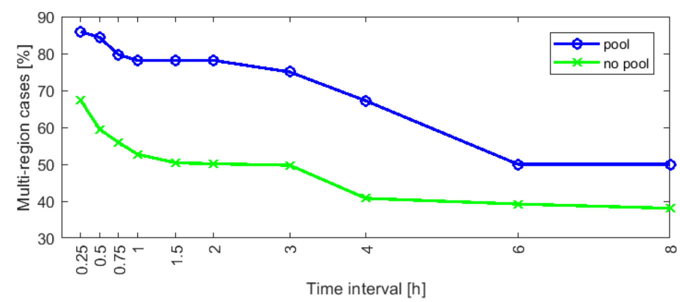


Fig. 6. The proportion of households where the multi-region case applies decreases with increasing time interval. The dependency holds for both households with and without swimming pool.

the dependency of the precision on the time interval. For all intervals the precision is in the range between 45% and 75% which is considerably higher than the precision of the all-positive classifier. Although the values change in a range of about 30%, due to the small number of positives (64) these differences are not statistically significant.

While the precision shows no clear tendency, Figure 5 illustrates that an increase of the time interval negatively affects the number of detected pools (recall). The reason for this behavior is mainly the effect that with lower time resolution, more single-region cases (hence fewer multi-region cases) occur (Figure 5). Since single-region cases are classified as having no swimming pool (Section III-B), the recall decreases.

Figure 6 shows that multi-region cases are converted to single-region cases in general, i.e., it occurs for both households with and without swimming pools.

Although our analysis method is very different from the method in [13] the results obtained in this paper are analogous to theirs: with decreasing time granularity the recall drops while the precision remains rather constant.

V. Conclusion

We showed that it is possible to detect swimming pools in 15-minute household load data by relatively simple algorithms. The privacy-relevant precision achieved by our classifier is nine times higher than that of an uninformed classifier. Due to the correlation between swimming pools and more expensive homes, the ability to detect swimming pools can be considered a privacy issue. Since the European Union requires smart meters to provide measurement intervals not longer than 15 minutes, the performance of our algorithm suggests that a 15-minute interval might still be too short from a privacy perspective. The proposed approach is surprisingly robust and works at even lower resolutions in the sense that the precision remains high. We showed that in turn using coarser resolutions would enhance privacy by decreasing the recall, i.e., number of detected households with swimming pools.

Acknowledgment

The financial support by the Austrian Federal Ministry of Science, Research and Economy, the Austrian National Foundation for Research, Technology and Development and the Federal State of Salzburg is gratefully acknowledged. Futhermore, the authors would like to thank the *Energieinstitut* at the Johannes Kepler University Linz for providing the data set.

References

[1] M. R. Asghar, G. Dán, D. Miorandi, and I. Chlamtac, "Smart Meter Data Privacy: A Survey," *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2820–2835, 2017.

[2] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin, "Private memoirs of a smart meter," in *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, ser. BuildSys '10. New York, NY, USA: ACM, 2010, pp. 61–66. [Online]. Available: http://doi.acm.org/10.1145/1878431.1878446

[3] D. C. Bergman, D. Jin, J. Juen, N. Tanaka, C. Gunter, and A. Wright, "Distributed Non-Intrusive Load Monitoring," in *Proceedings of the IEEE/PES Conference on Innovative Smart Grid Technologies (ISGT 2011), Anaheim, CA, USA, January 2011*, 2011.

[4] J. Kolter and M. J. Johnson, "Redd: A Public Data Set for Energy Disaggregation Research," in *Workshop on Data Mining Applications in Sustainability (SIGKDD)*, aug 2011, pp. 1–6. [Online]. Available: http://redd.csail.mit.edu/kolter-kddsust11.pdf

[5] M. Zeifman and K. Roth, "Nonintrusive Appliance Load Monitoring: Review and Outlook," *IEEE Transactions on Consumer Electronics*, vol. 57, pp. 76–84, 2011.

[6] A. Zoha, A. Gluhak, M. A. Imran, and S. Rajasegarar, "Non-intrusive Load Monitoring approaches for disaggregated energy sensing: A survey," *Sensors (Switzerland)*, vol. 12, no. 12, pp. 16 838–16 866, 2012.

[7] European Commission, "2012/148/EU: Commission Recommendation of 9 March 2012 on preparations for the roll-out of smart metering systems," 2012. [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32012H0148

[8] M. a. Lisovich and S. B. Wicker, "Privacy Concerns in Upcoming Residential and Commercial Demand-Response Systems," *IEEE Proceedings on Power Systems*, vol. 1, no. 1, pp. 1–10, 2008.

[9] U. Greveler, B. Justus, and D. Löhr, "Multimedia Content Identification Through Smart Meter Power Usage Profiles," in *Proceedings of the 2012 International Conference on Information and Knowledge Engineering (IKE'12)*, Las Vegas, USA, 2012.

[10] H. Kim, M. Marwah, M. F. Arlitt, G. Lyon, and J. Han, "Unsupervised Disaggregation of Low Frequency Power Measurements," in *The 11th SIAM International Conference on Data Mining*, 2011, pp. 747–758.

[11] J. Z. Kolter and T. Jaakkola, "Approximate Inference in Additive Factorial HMMs with Application to Energy Disaggregation," *Journal of Machine Learning Research - Proceedings Track*, vol. 22, pp. 1472–1482, apr 2012.

[12] G. W. Hart, "Nonintrusive Appliance Load Monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[13] G. Eibl and D. Engel, "Influence of Data Granularity on Smart Meter Privacy," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 930–939, 2015.

[14] D. Chen, S. Barker, A. Subbaswamy, D. Irwin, and P. Shenoy, "Non-Intrusive Occupancy Monitoring using Smart Meters," in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings - BuildSys'13*, 2013, pp. 1–8.

[15] G. Eibl, S. Burkhart, and D. Engel, "Unsupervised Holiday Detection from Low-Resolution Smart Metering Data," in *Proceedings of the 4th International Conference on Information Systems Security and Privacy, ICISSP 2018*. SciTePress, 2018, p. to appear.

[16] N. Pathak, D. Lachut, N. Roy, N. Banerjee, and R. Robucci, "Non-Intrusive Air Leakage Detection in Residential Homes," in *ICDCN '18: 19th International Conference on Distributed Computing and Networking*. Varanasi, India: ACM, 2018.

[17] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. Pearson, 2017.

[18] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.