Spatial Context Awareness in Surgery through Sound Source Localization

Matthias Seibold[1], Ali Bahari Malayeri[1], and Philipp Fürnstahl[1,2]

[1] Research in Orthopedic Computer Science (ROCS), Balgrist University Hospital, University of Zurich, Zurich, CH-8008, Switzerland
[2] OR-X Translational Center for Surgery, Balgrist University Hospital, University of Zurich, Zürich, Switzerland

Abstract. Context awareness and scene understanding is an integral component for the development of intelligent systems in computer-aided and robotic surgery. While most systems primarily utilize visual data for scene understanding, recent proof-of-concepts have showcased the potential of acoustic signals for the detection and analysis of surgical activity that is associated with typical noise emissions. However, acoustic approaches have not yet been effectively employed for localization tasks in surgery, which are crucial to obtain a comprehensive understanding of a scene. In this work, we introduce the novel concept of Sound Source Localization (SSL) for surgery which can reveal acoustic activity and its location in the surgical field, therefore providing insight into the interactions of surgical staff with the patient and medical equipment.
We show the potential of this concept by interpreting sound activity heatmaps using an acoustic camera in two proof-of-concept localization tasks, an object detection task for surgical sawing and a keypoint detection task for surgical chiseling. We achieve an AP at 0.5 IoU of 86.07% for the object detection task and a mean euclidean distance of 13.70 ± 14.65 px at an image resolution of 1100x825 px for the keypoint detection task. Based on these results, we believe that the localization of acoustic events has great potential for surgical scene understanding, opening up many new research directions for multimodal sensing solutions in the operating room of the future. To the best knowledge of the authors this is the first work that proposes to leverage SSL in the medical context.

Keywords: Computer Aided Surgery · Surgical Context Awareness · Sound Source Localization · Acoustic Sensing

1 Introduction

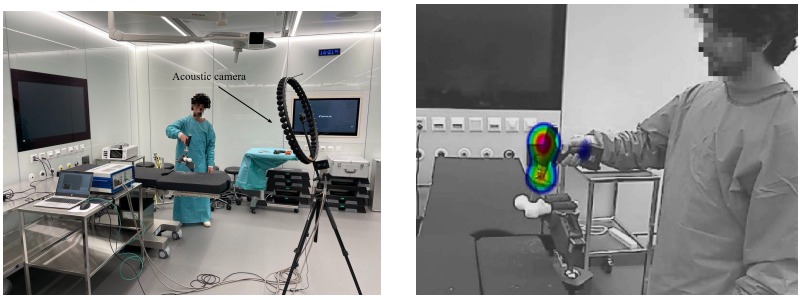
Surgical interventions involve complex interactions between surgical staff, medical devices, instruments, and patients. As the basis for the next generation of computer aided surgery systems, the modelling of events, structure, and interactions in the operating room is a crucial component to enable complex scene understanding for intelligent systems [7].

In previous work, many efforts have been made to describe surgical interventions in a systematic way. Surgical Process Modeling (SPM) has been introduced to provide a simplified structural representation of surgeries by decomposing the surgical process into its main phases, individual steps and sub-steps [8]. The field of surgical workflow recognition deals with the automated analysis and detection of high-level surgical phases based on sensor data captured in the operating room. Surgical workflow recognition can be employed to provide feedback to the surgical staff, trigger alarms in the case of adverse events, or facilitate the analysis of surgeries for documentation and surgical training. Optical data from endoscopic video or external camera views captures the activity in the surgical field and is the modality of choice for most surgical workflow recognition systems [11].

As an extension to high-level surgical phase labels, action triplets were proposed to create a more fine-grain representation of surgical activities in the operating room [9]. More recently, methods have been proposed to explicitly model the links and interactions of surgical staff, patients, and medical devices in the operating room as a graph. The methods dynamically build a graph structure, incorporating actors (surgeons, patients) and objects (medical instruments, devices, and equipment) as nodes and relationships and interactions between them as edges. The graph representation can subsequently be used for downstream tasks, such as clinical role prediction [10] or activity recognition [6]. Many works hereby rely on spatial information to generate the surgical context, as the geometric arrangement of surgical staff, instruments, medical devices, and patients contains highly relevant information about surgical procedures [10,16,6]. In this context, the need for expensive localization labels is a limiting factor for the generation of large-scale datasets [17].

Recently, acoustic signals have been identified to have great potential for the development of novel multimodal sensing solutions in the medical context. Using Acoustic Sensing (AS), many surgical activities which are associated with the emission of typical acoustic signals, such as coagulation, suction, sawing, milling, drilling, etc. can be detected for surgical context analysis [13]. Furthermore, AS can be utilized to detect short-time events in the surgical context, e.g. drill breakthrough in bone drilling [15] or the identification of different states in surgical milling [1], and can be utilized where vision-based systems are impractical or measurements are not feasible to obtain with optical sensors, e.g. for the monitoring of the implant insertion process in orthopedic surgery [4,14]. While AS systems showed promise, current systems are not able to determine the spatial locations of acoustic events in the operating room. The localization of acoustic events related to surgical activity in the operating field would enable spatial context awareness with activity- and context-related information for intelligent systems in surgery.

In this work, we propose a novel concept to enable spatial context awareness of surgical activity by introducing SSL for surgical procedures. While this is a novel concept for medical applications, SSL has been investigated in prior work for a variety of applications, e.g. noise source identification in industrial



(a) Experimental Setup (b) Acoustic Image

Fig. 1: **a)** The experimental setup described in section 2.1, including the circular phased microphone array on the right. **b)** An acoustic image overlaid onto a greyscale optical image captured from our experimental setup during sawing with a surgical power tool (Arthrex DrillSaw Max 600 (Arthrex Inc., 1370 Creekside Blvd., Naples, Florida 34108)). The measurement reveals two distinct sound source locations from the saw blade connector and the motor. The colors encode the measured sound pressure levels with a colormap ranging from red to blue from high to low sound pressure.

applications [2], the localization of emergency vehicles for autonomous driving [3], or in the context of robot audition [12]. To reveal acoustic activity and its location in the surgical field, we employ an *Acoustic Camera*, a system consisting of a phased microphone array with known geometry that is able to compute an acoustic image of a scene that contains the acoustic signal power for each pixel. The obtained SSL heatmap can be calibrated and overlaid onto an optical image to reveal the visual context of the acoustic activity. We have verified the applicability of this new concept for describing the surgical context in two first applications, an object detection task for surgical sawing, and a keypoint detection task for the identification of the entry point in surgical chiseling. We believe that SSL as an extension of acoustic sensing concepts can play an important role for advancing the context-awareness of intelligent systems for the surgery of tomorrow e.g. for surgical robots that perform tasks autonomously or collaborate with the surgical staff. Here, the localization of surgical sound events can help to create a better internal digital representation of the world and enable these systems to better understand their surrounding space. The associated code and data has been made publicly available under https://rocs.balgrist.ch/en/open-access/.

2 Materials and Method

2.1 Experimental Setup and Sound Source Localization

To assess the potential of SSL in the context of surgical interventions, we created an experimental setup in a real operating room, consisting of a 3d-printed model

of a human femur segmented from CT data which was fixated to a surgical operating table. Two types of surgical actions common in orthopedic surgery were performed, namely sawing and chiseling. To reveal the acoustic activity and its locations within the surgical field, we utilize a commercially available Acoustic Camera, the gfai Ring48 (gfai tech GmbH, Volmerstraße 3, 12489 Berlin, Germany), consisting of a circular array of $M = 48$ microphones and a calibrated optical camera, as illustrated in figure 1 a).

Figure 1 b) shows an example acoustic image from surgical sawing overlaid onto the calibrated optical image. This heatmap is computed by time-domain beamforming (*delay-and-sum beamforming*), defined by equation 1. The resulting sound intensity $L(t, x_i)$ is evaluated for every pixel of the search grid defined at focus distance $d = x - x_i$ from the microphone array. Hereby, p_m corresponds to the signal acquired from each microphone, x is the microphone location and x_i is expected source location corresponding to a pixel in the search grid. The time signal measured by each microphone is delayed by the respective retarded time $t_0 = |x - x_i|/c_0$, where c_0 represents the speed of sound in the medium.

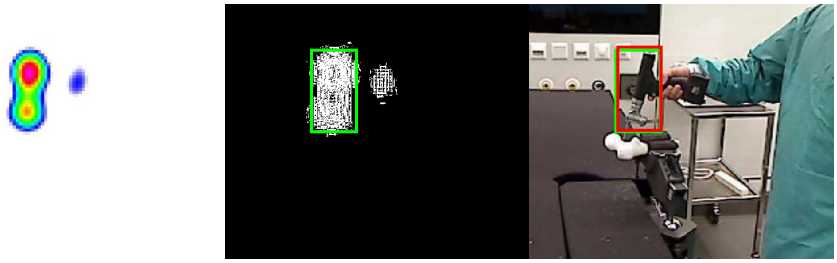
$$L(t, x_i) = \frac{4\pi}{M} \sum_{m=1}^{M} p_m(x_i, t + t_0)|x - x_i| \tag{1}$$

2.2 SSL-based Object Detection for Surgical Sawing

Object detection is a fundamental task in computer vision for detecting object instances in images which is utilized in surgical applications for example to detect surgical tools and their locations within the surgical field. As a first application, we investigated the potential of SSL for solving the object detection task in surgical bone sawing which is an integral part of many surgical procedures.

We recorded 201 sawing procedures in different locations of the bone phantom and with different sawing angles. Samples were obtained with a length of 1 s and a sampling rate of 192 kHz. We varied the distance $d = [1.2m, 2m]$ from the sensor to the setup, adjusted the focus distance, respectively, and randomized the viewing direction of the acoustic camera to the area of interest. To compute the SSL heatmap, we empirically identified optimal parameters and applied a band-pass filter from 4.5 kHz to 11.5 kHz while restricting the dynamic range of the resulting acoustic image to 18.7 mPa of sound pressure. Recording and preprocessing was executed using the professional sound analysis software NoiseImage 4.13.0 (gfai tech GmbH, Volmerstraße 3, 12489 Berlin, Germany). All sound source location maps were computed with a fixed size of 200x200 pixels.

Exploration of the data showed that the surgical sawing activity is spatially represented by two distinct sound sources on the drill body, as illustrated in figure 2. Subsequently, we apply image processing techniques, namely color filtering and connected component analysis to the heatmap to compute the bounding box estimates for the measured acoustic patterns. Finally, we select the largest detected bounding box as the prediction for the main body of the drill. We manually labeled the ground truth bounding box of the surgical saw in the

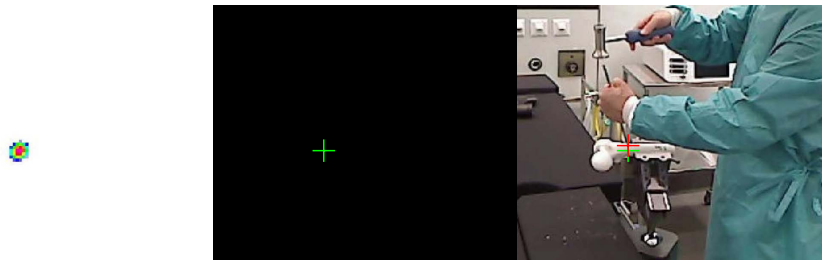


(a) SSL heatmap (b) Filtered and Prediction (c) Result vs. Ground Truth

Fig. 2: An illustration of the pipeline for SSL-based object detection for surgical sawing. **a)** The SSL heatmap contains the measured sound activity and locations. **b)** The acoustic image is color filtered and region proposals are computed based on connected component analysis, the largest detecting bounding box is selected as the final prediction. **c)** The calibrated color image with predicted (green) and ground truth (red) bounding boxes overlaid.

optical images as reference labels. Associated code and data will be made publicly available upon acceptance.

2.3 SSL-based Keypoint Detection for Surgical Chiseling



(a) SSL heatmap (b) Filtered and Prediction (c) Result vs. Ground Truth

Fig. 3: An illustration of the pipeline for SSL-based keypoint detection in surgical chiseling with **a)** the raw SSL heatmap, **b)** the filtered acoustic images with detected centroid, and **c)** the calibrated color image with predicted (green) and ground truth (red) keypoint overlaid.

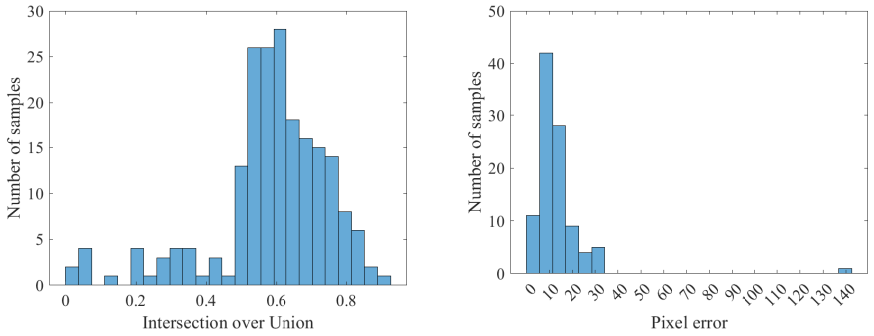
A second promising direction is the detection of contact points between tool and anatomy, in this example of a surgical chisel and the bone phantom during surgical chiseling. For systematic analysis, we captured a total number of 100 chiseling actions, and manually segmented each individual hammer blow from

the recordings, resulting in clips of 0.5 - 1 s. The screening of the data revealed that the contact point between chisel and anatomy can be derived from the propagation of the structure-borne vibrations induced by the hammer blow and chisel into the bone. We restrict the dynamic range of the resulting acoustic image to 52 mPa of sound pressure and do not apply any further filtering in the frequency domain for computing the final SSL heatmaps.

We export the acoustic image for each measurement, apply color filtering, and connected component analysis, as described in section 2.2, and compute the keypoint as the centroid of the largest detected component. We manually labeled the ground truth keypoint as the contact point of chisel and bone phantom in the optical images as reference labels. This process is illustrated in figure 3.

3 Results and Evaluation

We separately evaluate both tasks presented in section 2 and compute the respective relevant metrics over the entire data collected in each experiment, respectively. For assessing the performance of bounding box detection in surgical sawing, we compute the performance metric as the Average Precision (AP) at 0.5 Intersection over Union (IoU), where our proposed method achieves an AP at 0.5 IoU of 86.07%. For the tasks of keypoint detection in surgical chiseling, we compute the Euclidean distance from the predicted point to the ground truth point in pixels (px). Our method achieves an average error of 13.70 ± 14.65 px with a Median of 11.18 px at an image resolution of 1100 x 825 px.



(a) Object Detection (b) Keypoint Detection

Fig. 4: Result distributions for the two presented applications, **a)** Bounding Box Detection in Surgical Sawing and **b)** keypoint detection in surgical chiseling.

In figure 4, we present the result distribution for both presented tasks to illustrate the performance over all samples in the respective datasets.

4 Discussion

In this work, we introduce the new concept of SSL in surgery for acoustic-based scene understanding and context description and evaluate this novel concept in two exemplary applications. While our work is preliminary, we show that SSL holds significant promise as a complementary approach to visual systems, enhancing surgical context awareness in selected indications. We envision SSL as an important foundation that could lay the groundwork for novel methodologies in surgical scene understanding, pose estimation of surgical tools, and surgical data science.

While the majority of the results show good performance for both tasks, the presence of outliers is apparent, as can be observed in figure 4. These outliers can be attributed to the fact that the measurement equipment is very sensitive and captures all acoustic reflections from the room and other objects in the field of interest. To gain additional insights on the application in the field, the experiments were intentionally conducted in a real surgery room consisting of audio-reflective glass walls without acoustic treatment. As a result, reflections and standing waves can appear as artifacts in the computed SSL heatmap. Sometimes, these artifacts appear with the highest amplitude which results in false location predictions by the proposed algorithm, which can be observed in figure 4. While outliers are scarce, this phenomenon should be investigated in further research to increase the robustness of the proposed method. A promising direction to address this problem is the application of deep learning-based SSL approaches which have shown superior performance, especially for indoor environments with reverberation and diffuse noise [5]. Furthermore, the number of microphones and geometry of the microphone array has an important influence on the resolution and accuracy of the computed heatmaps which should be optimized in future research.

The generation of the acoustic image is an expensive operation, as it requires an individual computation of the beamforming algorithm for each pixel, respectively. While the operation is highly parallelizable on a GPU ($\sim$2 seconds on a NVIDIA RTX 2080 SUPER for an acoustic image with size 200x200), the computational cost scales up quadratically with increased image resolution. In this context, learning-based methods will be necessary in the future to speed up the computation and achieve the real-time performance required for intraoperative applications.

A limitation for SSL technology in the context of surgical scene understanding is that it only allows to detect surgical events that are associated with specific acoustic signal emission. For instance, object detection is only possible for parts that emit sounds which required us to exclude the handle of the tool from the bounding box detection. While we identified great potential for orthopedic surgery, the applicability to other surgical fields needs to be investigated in future work. In this context, an example for a non-orthopedic use case would be the localization of coagulation and suction events.

To obtain suitable heatmaps, the system parameters of the acoustic camera were empirically calibrated to the respective tool, i.e. the dynamic range and

filter parameters. In future research, these parameters need to be systematically analyzed to achieve generalization to multiple surgical scenarios. As an extension to detection a tool as a whole, the optimization of parameters could allow the detection of two or more distinct parts of a surgical tool which could allow to deduce even an estimate for the orientation of the tool which should be investigated in future research.

While we present only a proof-of-concept using a single 3d-printed bone phantom, we believe that our experimental setup is realistic enough to illustrate the potential of SSL for context understanding and can be easily transferred to more realistic contexts. To account for a realistic acoustic environment, we performed all experiments in a real surgery room, including air ventilation and realistic surgical equipment, i.e. a real surgery table, instruments, and clothing.

Even though it has been shown, that scene graphs can be generated without location information [17], the localization of acoustic events in surgery contains valuable information for many downstream tasks. As our approach does not require any additional manual labeling, the location information can be obtained inexpensively. The current implementation provides a proof-of-concept for two-dimensional location detection which we want to extend into three-dimensional space in future work. This will allow to fuse location information of acoustic events with other 3d data, i.e. human and object pose estimates or semantic labels for scene description, and obtain a fine-grain 3d description of a surgical scene. As the first work to introduce SSL in the surgical context, we believe that this paper will enable the field of computer aided surgery to further leverage multimodal data in the surgical context.

The proposed concept of using SSL for surgical context understanding opens up many promising and interesting research directions and algorithms to be developed to improve the performance and accuracy of the technology, like artifact reduction and the optimization for real-time. Here, learning-based methods will have great potential for processing, optimizing sound activity heatmaps and fusing them with visual data for improved surgical context understanding.

5 Conclusion

In this work, we propose to extend acoustic sensing in surgery with spatial information through SSL to enable spatial context awareness of acoustic activity in the operating field. We showcase the potential of SSL in an experiment, employing beamforming with a commercial acoustic camera to determine the location of sound sources during relevant surgical activities. We formulate two tasks, an object detection task for surgical sawing, and a keypoint detection task for surgical chiseling for which our method achieved promising performance in accurately determining the spatial location of the surgical activity in the operating field. The presented concept opens up many new research directions for the usage of acoustic signals in surgical scene understanding and can be an important contribution for the development of intelligent systems in the operating room of the future.

Acknowledgments. This work has been supported by the OR-X - a swiss national research infrastructure for translational surgery and associated funding by the University Hospital Balgrist. We furthermore thank Dr. Matthias Brechbühl from Norsonic Brechbühl AG and Carsten Hessenius from gfai GmbH for the support that made this work possible.

Disclosure of Interests. All authors declare that they have no conflicts of interest.

References

1. Dai, Y., Xue, Y., Zhang, J.: State identification based on sound analysis during surgical milling process. In: 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO) (2015)
2. Fedorko, G., Liptai, P., Molnár, V.: Proposal of the methodology for noise sources identification and analysis of continuous transport systems using an acoustic camera. Engineering Failure Analysis **83**, 30–46 (2018)
3. Furletov, Y., Willert, V., Adamy, J.: Auditory scene understanding for autonomous driving. In: 2021 IEEE Intelligent Vehicles Symposium (IV). pp. 697–702 (2021)
4. Goossens, Q., Pastrav, L., Roosen, J., Mulier, M., Desmet, W., Vander Sloten, J., Denis, K.: Acoustic analysis to monitor implant seating and early detect fractures in cementless tha: An in vivo study. Journal of Orthopedic Research (2020)
5. Grumiaux, P.A., Kitić, S., Girin, L., Guérin, A.: A survey of sound source localization with deep learning methods. The Journal of the Acoustical Society of America **152**(1), 107–151 (2022)
6. Hamoud, I., Jamal, M.A., Srivastav, V., MUTTER, D., Padoy, N., Mohareri, O.: St(or)^2: Spatio-temporal object level reasoning for activity recognition in the operating room. In: Medical Imaging with Deep Learning. vol. 227, pp. 1254–1268 (2024)
7. Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., Hashizume, M., Katic, D., Kenngott, H., Kranzfelder, M., Malpani, A., März, K., Neumuth, T., Padoy, N., Pugh, C., Schoch, N., Stoyanov, D., Taylor, R., Wagner, M., Hager, G.D., Jannin, P.: Surgical data science for next-generation interventions. Nature Biomedical Engineering **1**, 691–696 (2017)
8. Neumuth, T.: Surgical process modeling. Innovative Surgical Sciences **2**(3), 123–137 (2017)
9. Nwoye, C.I., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 364–374 (2020)
10. Özsoy, E., Örnek, E.P., Eck, U., Czempiel, T., Tombari, F., Navab, N.: 4d-or: Semantic scene graphs for or domain modeling. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 475–485 (2022)
11. Padoy, N.: Machine and deep learning for workflow recognition during surgery. Minimally Invasive Therapy & Allied Technologies **28**(2), 82–90 (2019)
12. Rascon, C., Meza, I.: Localization of sound sources in robotics: A review. Robotics and Autonomous Systems **96**, 184–210 (2017)
13. Seibold, M., Hoch, A., Farshad, M., Navab, N., Fürnstahl, P.: Conditional generative data augmentation for clinical audio datasets. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 345–354 (2022)
14. Seibold, M., Hoch, A., Suter, D., Farshad, M., Zingg, P.O., Navab, N., Fürnstahl, P.: Acoustic-based spatio-temporal learning for press-fit evaluation of femoral stem implants. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 447–456 (2021)
15. Seibold, M., Maurer, S., Hoch, A., Zingg, P., Farshad, M., Navab, N., Fürnstahl, P.: Real-time acoustic sensing and artificial intelligence for error prevention in orthopedic surgery. Scientific Reports **11** (2021)

16. Özsoy, E., Czempiel, T., Holm, F., Pellegrini, C., Navab, N.: Labrad-or: Lightweight memory scene graphs for accurate bimodal reasoning in dynamic operating rooms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2023)
17. Özsoy, E., Holm, F., Czempiel, T., Navab, N., Busam, B.: Location-free scene graph generation. arXiv:2303.10944 (2023)