Relevance-Based Visualization to Improve Surgeon Perception

Olivier Pauly[1], Benoît Diotte[1], Séverine Habert[1], Simon Weidert[2], Ekkehard Euler[2], Pascal Fallavollita[1,⋆], and Nassir Navab[1]

[1] Chair for Computer Aided Medical Procedures, Technische Univ. München, Germany
[2] Chirurgische Klinik und Poliklinik Innenstadt, München, Germany
fallavol@in.tum.de

Abstract. In computer-aided interventions, the visual feedback of the doctor is vital. Enhancing the relevant object will help for the perception of this feedback. In this paper, we present a learning-based labeling of the surgical scene using a depth camera (comprised of RGB and depth range sensors). The depth sensor is used for background extraction and Random Forests are used for segmenting color images. The end result is a labeled scene consisting of surgeon hands, surgical instruments and background labels. We evaluated the method by conducting 10 simulated surgeries with 5 clinicians and demonstrated that the approach provides surgeons a dissected surgical scene, enhanced visualization, and upgraded depth perception.

Keywords: visualization, medical augmented reality, machine learning, multimodal image fusion, operating room.

1 Introduction

Humans boast a sophisticated cognitive system which takes approximately 15-20 different psychological stimuli into account in order to perceive spatial relationships between objects [1]. Nevertheless, in complex settings, such as the operating room theatre, the cognitive system is challenged as clinicians are confronted with information stemming from multiple sources when making surgical decisions. Presenting all of the information in an effective manner is a difficult task. Consequently, improving the understanding and perception of clinicians towards their surgical environment becomes an important feedback for the success of computer-assisted intervention applications (e.g. labeling the surgeons action helps in workflow analysis [2], or improving surgeon visualization of fused modalities helps successful patient outcomes).

This feedback can be provided by mixed and augmented reality (AR) visualizations for use in computer-assisted interventions. However, few of these systems have been introduced for daily use into the operating room (OR). This may be

⋆ Corresponding author.

D. Stoyanov et al. (Eds.): IPCAI 2014, LNCS 8498, pp. 178–185, 2014.

the result of several factors: the systems are developed from a technical perspective, are rarely evaluated in the field, and/or lack consideration of the clinician and the constraints of the OR [3].

As of late, the community did achieve success in deploying the first medical augmented reality technology (an AR mobile fluoroscope) within orthopedic and trauma surgery rooms, and this recent introduction promises to support surgeons in their understanding of the spatial relationships between anatomy, implants and their surgical tools [4,5]. The output overlay of such a technology is a uniform alpha-blending between the X-ray and optical images. The issue with this blending type is that the understanding of the scene can be altered when the field of view of the scene becomes highly cluttered (e.g. with surgical tools and implants). It becomes increasingly difficult to rapidly recognize and differentiate different structures in the fused image. Moreover, the clinicians depth perception is altered as (i) the X-ray anatomy appears floating on top of the scene in the optical image, (ii) hands and surgical instruments occlude the visualization, and (iii) there is no correct ordering between structures in the fused images.

With these issues in mind, we note that all pixels in X-ray and optical images do not have the same importance and contribution to the final blending (e.g. the background is not important compared to the surgical tool). This observation suggests extracting only relevant-based data according to pixels belonging to background, tools and clinician hands [6]. The labeling of the surgical scene by a precise segmentation and differentiation of its different parts allows a relevant blending respecting the desired ordering of structures. A few attempts have been endeavored, such as in [7]. In these early works, a Naive Bayes classification approach based on color and radiodensity is applied to recognize the different objects in X-ray and color images. Depending on the pair of pixels it belongs to, each pixel is associated to a mixing value to create a relevant-based fused image. While authors showed promising results, recognizing each object on their color distribution is very challenging and not robust to changes in illumination.

Contribution: We introduce a surgical scene labeling paradigm based on machine learning and having as input both an optical and depth camera in a medical AR setting. In our application, the depth is a useful hint for the segmentation and ordering of hands and tools with respect to anatomy since the clinician performs surgery over the patient. Thus, our visualization paradigm is founded on segmentation consisting in modeling the background via depth data. We perform in parallel color image segmentation via the state-of-the-art Random Forests. To refine our segmentation method we use the GrabCut algorithm. Lastly, we combine our background modeling and color segmentation in order to identify the objects of interests in the color images and achieve successfully ordering of structures. We conducted 10 simulated surgeries with 5 clinicians to showcase our visualization results.

2 Methods

A depth camera with an integrated optical camera (Asus Xtion Pro Live) is affixed to a mobile C-arm fluoroscope above of the surgery workspace, giving a general overview of clinician gestures and surgical tool manipulations. The depth camera is positioned at its fabricated optimal visual focal length (70cm) of the patient table. Since the camera has a visual view of the surgical scene it is reasonable to assume that the hands and surgical instruments are on top of or at the same level as the patient. The depth image is built-in registered to the RGB camera therefore the image I and depth image D are defined on the same domain $\Omega \in \mathbb{R}^2$ with I and D being defined respectively as $I : \sigma \rightarrow \mathbb{R}^3$ and $D : \sigma \rightarrow \mathbb{R}$.

2.1 Identifying Objects of Interest in RGB-D

The objective is to dissect the surgical scene using the images from the RGB and depth camera. We divide the scene into 3 classes $C = \{tool, hands, background\}$. The surgeon actions via tools and hands are combined to form the foreground class (closer to the camera). We use the depth image to create a background model that will, for every frame, give a probability at a given pixel x, $P_D(f^c|x)$ of belonging to the background (f^c, complement class of the foreground). With the RGB images, the probabilities $P_I(c|x)$ of belonging to the tools, the surgeon hand or the background is obtained using Random Forests. Then, since the modalities RGB and depth are independent (the color is not interfering on the depth), we can decompose the joint distribution of a pixel belonging to the foreground and to an object c $P_{I,D}(f, c|x)$ as

$$P_{I,D}(f, c|x) = (1 - P_D(f^c|x))P_I(c|x) \tag{1}$$

Background Extraction Using Depth Images. Background modeling has been widely studied for performing background subtraction in color images in tracking applications. In a fixed camera setup, the key idea is to learn a color distribution for each pixel from a set of background images. As reported in [8], several approaches have been proposed within the last decade for adaptive real-time background subtraction based on running Gaussian averages, mixture models, kernel density estimation or the so-called Eigenbackground. In the present work, we propose to learn a fixed background model using the depth image and based solely on a set of acquired depth frames at the beginning of the surgery. A fixed model is more suitable to our application since adaptive models presume that foreground objects are moving fast, while in surgery, the object of interest (hands or tools) may stay immobile the majority of the time. Formally, we consider a set of N depth frames D accumulated at the beginning of the surgical sequence, when no objects of interest are present in the scene. We consider the background model at each pixel $x \in \Omega$ as a univariate gaussian model where the mean and variance of this distribution are the values measured over the set of

frames D at the pixel $x \in \Omega$. Lastly, in the remaining images of the sequence (objects of interests enter the scene), a background probability image is created for each individual frame.

Segmentation by Random Forest of RGB Images. As reported in [9], random forests have found a wide variety of applications in medical image analysis such as anatomy localization, segmentation or lesion detection. As an ensemble of decision trees, they provide piecewise approximations of any distribution in high-dimensional space. In our case, we model the probability $P_I(c|x)$ $x \in \Omega$ to belong to a class $c \in C = \{tool, hands, background\}$. The visual content of a pixel x is defined by a feature vector $\mathbf{X} \in \mathbb{R}^d$. $\mathbf{X}$ encodes the mean intensity value computed in d rectangular regions of different sizes in the neighborhood of x in the color channels of the CIELab color space. Following a "divide" and "conquer" strategy, each tree t, $t \in \{1, T\}$, first partitions the feature space in a hierarchical fashion and then estimates the posterior probabilities in each "cell" of this space. Given a training set of pixels from different color images and their corresponding labels, a tree t aims at subdividing these data by using axis-aligned splits in $\mathbb{R}^d$ so that consistent subsets are created in its "leaves" in terms of their visual context and class information c. Each leaf of a tree models "locally" the posterior probability $P_I^t(c|x)$, encoded as a class histogram, computed from the set of observations reaching the leaf. At test time, the output of the trees can be combined by using posterior averaging: $P_I(c|x) = \sum_{t=1}^{T} P_I^t(c|x)$.

Object Extraction. For each frame, the joint probability can be calculated by multiplying the probability of belonging to the foreground $P_D(f|x)$ with the probability of belonging to any class c $P_I(c|x)$. Finally, the class label $\hat{c}$ of a pixel is estimated by finding the class whose probability $P_{I,D}(f, c|x)$ is higher, such as $\hat{c} = argmax_{c \in C} P_{I,D}(f, c|x)$.

Refinement Using GrabCut. Since the class estimations might be noisy, we choose to refine the current extraction of interest objects by a segmentation algorithm [10]. Known as GrabCut, this algorithm is an extension of the graph-cut framework that uses an efficient iterative estimation and handles incomplete labelling. GrabCut permits decreasing the labelling burden as it integrates 4 possible label classes: foreground, probably foreground, probably background, background. For more details, we refer the reader to [10]. In our case, to refine the extraction of tools in the frame, the pixels classified as tool by the Random Forest are labelled as possible foreground, the rest is labelled as background. GrabCut is then performed on the corresponding color frame using that labelling to provide a finer extraction of the tool. The same step is renewed also for the clinician's actions. Even though this step requires additional computations, it is fast and efficient, and permits to filter out some false positives or catch false negatives that comes from missing depth values. At this step, the different parts (background, clinician's hands and tools) have been classified in the image. This process is repeated for every frame of the video.

2.2 Application Using an Augmented Reality Fluoroscopy

Identifying Object of Interest in X-ray. We consider an X-ray image J that is co-registered to the color and depth images, with $J : \Omega \rightarrow \mathbb{R}$. To improve the alpha-blendings developed in [5,4], the segmentation in different clusters as previously described is used. However, to further improve the visualization, we also extract from the X-ray image J the objects of interest to the clinician (e.g. bones, implants). This classification task will assign a label $r \in \{0, 1\}$ for each pixel x, where $r = 1$ if x belongs to a relevant structure or $r = 0$ if not. In a probabilistic framework, we model the posterior distribution $P_J(r|x)$ by using a random forest. Similarly, the visual context of each pixel x is described by a feature vector $X \in \mathbb{R}^t$, encoding mean radiodensity values computed in t rectangular regions in its neighborhood. Once the forest has been trained by using a set of annotated images, a new incoming X-ray can be labelled by using $\hat{r} = argmax_{r \in \{0,1\}} P_J(r|x)$. Once the labelling is done, we refine with GrabCut the current segmentation. All the pixels classified as belonging as relevant structure ($r = 1$) are labeled as possible foreground and the rest is labeled as background.

Relevance-Based Image Fusion. The AR fluoroscopy technologies use an uniform alpha-blending to overlay the color images and the X-ray where the blending coefficient α is constant for all pixels. In this paper, we introduce a pixel dependant α parameter that changes values according to its belonging to an object of interest in the color image or in the X-ray image. Our new mixing paradigm is:

$$I_{overlay}(x) = \alpha(x)I(x) + (1 - \alpha(x))J'(x) \tag{2}$$

where J' is the 3-channels grayscale image corresponding to J such $J' = [J, J, J]$. Note that those values can be changed on the fly according to the will of the clinician, the type of clinician and also the different phases of the surgery workflow. For example, the value for the hands and tools can be decreased to allow the clinician to see the anatomy on the X-ray when performing distal locking on an intramedullary nail.

3 Experiments and Results

3.1 Evaluation of the Objects Identification

To evaluate the object identification algorithm for color images, 10 different orthopedic surgery simulations using a surgical phantom have been performed. Each simulation involves various clinician tasks and tools (clamps, screwdrivers, hammer, radiolucent drill, and scalpel). Each surgical simulation acquisition consisted of an average of 1000 frames. For the background modelling, the first 30 images ($\sim$ 1 second) of each sequence have been used to compute the background model.

Object Identification Using Color Images. In each of the 10 sequences, 4 video frames have been annotated. To describe the visual context of each pixel in the color image, 50 context features are extracted per CIElab channel. To tackle the task of object identification, a random forest classifier consisting in 20 trees of depth 15 is trained. After the first identification step, the GrabCut algorithm is executed using 2 iterations to refine the classifier results. The medium- to larger-sized surgical instruments are segmented very well. Minor segmentation errors occur specifically for the tip of the clamp allowing us to conclude that the segmentation algorithm needs further improvement to handle thin structures. The clinician's hands are globally well segmented over the various examples however we observe in some cases a wrong segmentation for the fingers primarily due to an aggressive GrabCut algorithm step that withdraws false positives, but also considers as background the pixels where the probability classes are too ambiguous to be considered as possible foreground. For quantitative results, we measure the accuracy of the classification into a class c thanks to the precision $\mathcal{P}$ and recall $\mathcal{R}$ measures over the annotated frames. We also calculate the DICE score $\mathcal{D}$, a similarity measure between the segmented class pixels and the annotated class pixels. The precision is over 0.8 for the hand, foreground and background classes, with a high score of 0.98 for the background, signifying that we have a good classification of most of the pixels belonging to those classes. Regarding the surgical tools, we achieve average precision with a value of 0.53 and a high standard deviation of 0.3. However, this global precision value can be decomposed to tool sizes as seen in Figure 3. As previously mentioned, medium to larger sized tools are generally well segmented. After further investigating our algorithm, the tools precision results can be explained by the amount of images used for the training of the Random Forest. Over the 38 training frames, each surgical tool appears in 5 images maximum and globally the presence of smaller tools in the training frames were much lower than the medium to larger sized tools. Resolving these issues will undoubtedly increase the precision values. Lastly, the recall values are really good for hand, foreground and background classes with values over 0.95, meaning that almost every annotated pixels have been recovered in those classes. The recall is good also for the surgical tool class. As a final note, a a clinician had their watch on and due to its black color, this structure was classified as a tool. In surgery and under sterile conditions, this issue would be resolved as the watch would be withdrawn. The computation time is 1.5 seconds per frame.

3.2 Fusion with X-ray Images

For the classification, 20 X-ray shots have been annotated. 50 context features are extracted by pixel, and the classifier consists in a random forest of 20 trees with depth 15. Then, the GrabCut algorithm is performed using 2 iterations to refine the segmentation of the object of interests.

Evaluation of Identification in X-ray Images. We use the same metrics (precision and recall) as with the RGB images. The recall and precision of both

background and foreground are close to 1, showing a good performance of the segmentation algorithm.

Evaluation of Fusion Results. The visualization results of the fused X-ray and RGB images are depicted in Figure 1. A qualitative evaluation of the relevance-based blending visualization compared to the uniform blending is performed. In total, 5 clinicians (3 experts surgeons and 2 last year medical students) provide their feedback using the traditional 5-pt Likert scale questionnaire (1- strongly disagree, 2-disagree, 3-neutral, 4-agree, and 5-strongly agree). Participants strongly agreed (4.6 ± 0.5) that the depth ordering is resolved using our approach (e.g. hands/tools first followed by patient/X-ray). Concerning the visibility of the instrument tip or the implants in X-ray, the feedback is respectively neutral (3.0 ± 1.4) and slightly positive (3.4 ± 1.1). Participants agreed

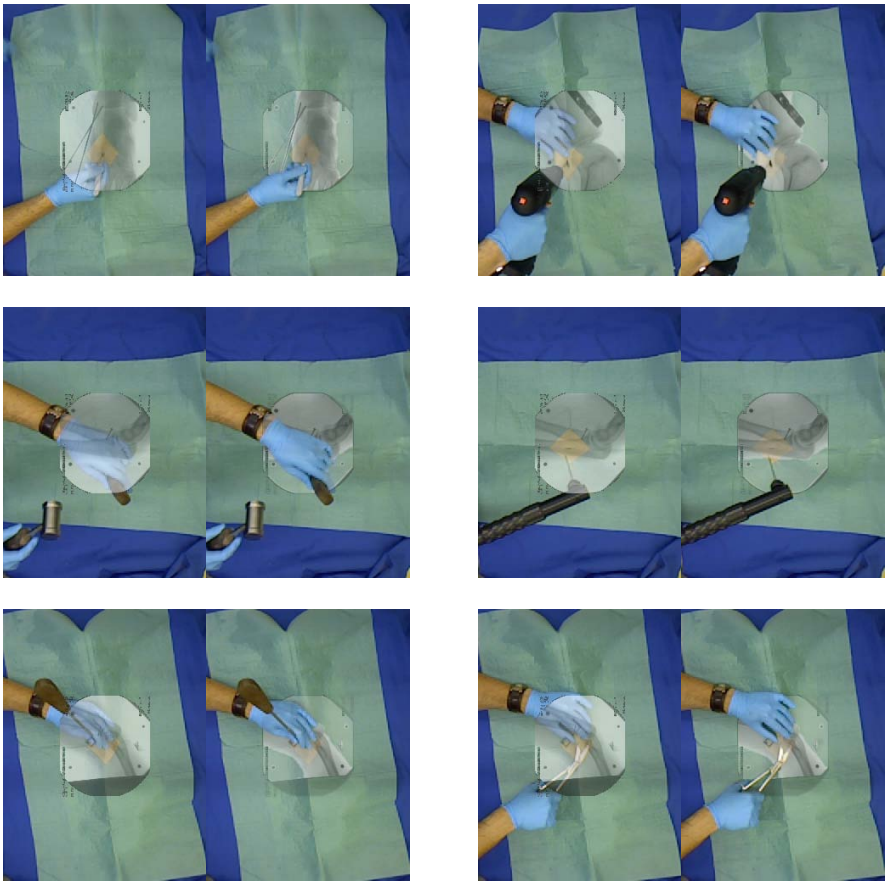

Fig. 1. Uniform and content-based visualizations over 6 frames

(4.0 ± 1.4) that the overall perception of the visualization is improved. Finally, all participants strongly agreed (4.6 ± 0.9) on the fact that they would prefer our new visualization compared to classical alpha blending found in the majority of registration algorithms in our community.

4 Conclusion

In this paper, we proposed a learning-based surgical scene labeling allowing the improved understanding and perception of various tasks when compared to the traditional alpha blending schemes. Our algorithm can detect the position and shape of the surgeon hands as well as the used tools. Our results are very promising for almost all objects, except smaller tools, but a more extended training phase should resolve this issue. We have demonstrated the applicability of our visualization framework in the context of existing medical augmented reality technologies. In future, our method can be extended to further applications such as 3D tool template matching, tool tracking and workflow analysis. Lastly, together with the IPCAI community, we hope to catalyze discussions on possible ways in improving visualization schemes that enable algorithms to "learn" what the surgeon wants to see during the surgical workflow phases.

References

1. Goldstein, B.E.: Sensation and perception. Cengage Learning (2013)
2. Padoy, N., Mateus, D., Weinland, D., Berger, M.O., Navab, N.: Workflow monitoring based on 3d motion features. In: IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), pp. 585–592. IEEE (2009)
3. Kersten-Oertel, M., Jannin, P., Collins, D.L.: Dvv: a taxonomy for mixed reality visualization in image guided surgery. IEEE Transactions on Visualization and Computer Graphics 18, 332–352 (2012)
4. Nicolau, S., Lee, P., Wu, H., Huang, M., Lukang, R., Soler, L., Marescaux, J.: Fusion of c-arm x-ray image on video view to reduce radiation exposure and improve orthopedic surgery planning: first in-vivo evaluation. In: 15th Annual Conference of the International Society for Computer Aided Surgery (2011)
5. Navab, N., Heining, S.M., Traub, J.: Camera augmented mobile c-arm (camc): Calibration, accuracy study, and clinical applications. IEEE Transactions on Medical Imaging 29, 1412–1423 (2010)
6. Pauly, O., Katouzian, A., Eslami, A., Fallavollita, P., Navab, N.: Supervised classification for customized intraoperative augmented reality visualization. In: IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 311–312 (2012)
7. Erat, O., Pauly, O., Weidert, S., Thaller, P., Euler, E., Mutschler, W., Navab, N., Fallavollita, P.: How a surgeon becomes superman by visualization of intelligently fused multi-modalities. In: SPIE Medical Imaging, pp. 86710L–86710L (2013)
8. Elgammal, A.: Background Subtraction: Theory and Practice. Springer (2013)
9. Criminisi, A., Shotton, J.: Decision Forests for Computer Vision and Medical Image Analysis. Springer (2013)
10. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics (TOG) 23, 309–314 (2004)