Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Machine learning-based augmented reality for improved surgical scene understanding

Olivier Pauly[a,c,*], Benoit Diotte[a], Pascal Fallavollita[a], Simon Weidert[b], Ekkehard Euler[b], Nassir Navab[a]

[a] *Computer Aided Medical Procedures, Technische Universität, München, Germany*
[b] *Chirurgische Klinik und Poliklinik Innenstadt, München, Germany*
[c] *Institute of Biomathematics and Biometry, Helmholtz Zentrum, München, Germany*

ARTICLE INFO

Article history:
Received 9 January 2014
Received in revised form 8 April 2014
Accepted 9 June 2014

ABSTRACT

In orthopedic and trauma surgery, AR technology can support surgeons in the challenging task of understanding the spatial relationships between the anatomy, the implants and their tools. In this context, we propose a novel augmented visualization of the surgical scene that mixes intelligently the different sources of information provided by a mobile C-arm combined with a Kinect RGB-Depth sensor. Therefore, we introduce a learning-based paradigm that aims at (1) identifying the relevant objects or anatomy in both Kinect and X-ray data, and (2) creating an object-specific pixel-wise alpha map that permits relevance-based fusion of the video and the X-ray images within one single view. In 12 simulated surgeries, we show very promising results aiming at providing for surgeons a better surgical scene understanding as well as an improved depth perception.

1. Introduction

In orthopedic and trauma surgery, the introduction of AR technology such as the camera augmented mobile C-arm promises to support surgeons in their understanding of the spatial relationships between anatomy, implants and their surgical tools [1,2]. By using an additional color camera mounted so that its optical center coincides with the X-ray source, the CamC system provides an augmented view created through the superimposition of X-ray and video images using alpha blending. In other words, the resulting image is a linear combination of the optical and the X-ray image by using the same mixing coefficient (alpha) over the whole image domain. While this embodies a simple and intuitive solution, the superimposition of additional X-ray information can harm the understanding of the scene for the surgeon when the field of view becomes highly cluttered (e.g. by surgical tools). It becomes more and more difficult to quickly recognize and differentiate structures in the overlaid image. Moreover, the depth perception of the surgeon is altered as the X-ray anatomy appears on top of the scene in the optical image.

In both the X-ray and the optical image, all pixels in the image domain do not have the same relevance for a good perception and understanding of the scene. Indeed, in the X-ray, while all pixels that belong to the patient bone and soft tissues have a high relevance for surgery, pixels belonging to the background do not provide any information. Concerning the optical images, it is crucial to recognize different objects interacting in the surgical scene, e.g. background, surgical tools or surgeon hands. First, this permits to improve the perception by preserving the natural occlusion clues when surgeon's hand or instruments occlude the augmented scene in the classical CamC view. Second, as a by-product, precious semantic information can be extracted for characterizing the activity performed by the surgeon or tracking the position of different objects present in the scene.

In this paper, we introduce a novel learning-based AR fusion approach aiming at improving surgical scene understanding and depth perception. Therefore, we propose to combine a mobile C-arm with a Kinect sensor, adding not only X-ray but also depth information into the augmented scene. Using the fact that structured light functions through a mirror, the Kinect sensor is integrated with a mirror system on a mobile C-arm, so that both color and depth cameras as well as the X-ray source have the same viewpoint. In this context of learning-based image fusion, a few attempts have been done in [3,4] based on color and X-ray information only. In these early works, a Naïve Bayes classification approach based on the color and radiodensity is applied to recognize the different objects in respectively the color and X-ray images from the CamC system. Depending on the pair of objects it belongs to,

[*] Corresponding author.
E-mail addresses: olivier.pauly@tum.de, pauly@cs.tum.edu (O. Pauly).

http://dx.doi.org/10.1016/j.compmedimag.2014.06.007
0895-6111/© 2014 Elsevier Ltd. All rights reserved.

each pixel is associated to a mixing value to create a relevance-based fused image. While this approach provided promising first results, recognizing each object on their color distribution only is very challenging and not robust to changes in illumination. In the present work, we propose to take advantage of additional depth information to provide an improved AR visualization: (i) we define a learning-based strategy based on color and depth information for identifying objects of interest in Kinect data, (ii) we use state-of-the-art random forest for identifying foreground objects in X-ray images and (iii) we use an object-specific mixing look-up table for creating a pixel-wise alpha map. In 12 simulated surgeries, we show that our fusion approach provides surgeons with a better surgical scene understanding as well as an improved depth perception.

2. Methods

2.1. System setup: Kinect augmented mobile C-arm

In this work, we propose to extend a common intraoperative mobile C-arm by mounting a Kinect sensor, that consists in a depth sensor coupled to a video camera. The video camera optical center of this RGB-D sensor is mounted so that it coincides with the X-ray projection center. The depth sensor is based on so-called structured light where infrared light patterns are projected into the scene. Using an infrared camera, the depth is inferred from the deformations of those patterns induced by the 3D structure of the scene. To register the depth images into the video camera coordinates, the sensor disposes of a built-in calibration. In the proposed setup illustrated by Fig. 2 on the left, the surgical scene is seen through a mirror system. Note that depth inference is still possible as the mirror perfectly reflects structured light without inducing deformations on the infrared patterns. Fig. 2 on the right shows our proof-of-concept setup we will use in our experiments. This system consists of an aluminium frame mimicing a C-arm with realistic dimensions, a Kinect sensor and a mirror system. Here we use one mirror to simulate the fact that the video optical center of the camera augmented mobile C-arm system has to virtually coincide with the X-ray source. While the effective range of the depth information is between 50 cm and 3 m, the distance between X-ray source and detector is about 1 m. The depth sensor is mounted at about 30 cm from the mirror, which is about 10 cm below the X-ray source. This effectively allows a depth range of about 70 cm from the detector. The fastest and highest synchronized resolution of the Kinect is 640×480 at 30 fps. Since the field of view of the Kinect is larger than the mirror, the images are cropped at 320×240 to fit the mirror view. In our experiments, real X-ray shots acquired from different orthopedic surgeries will be manually aligned into the view of our scene before starting our surgery simulations. In the next section, we will describe our novel learning-based AR visualization that combines intelligently these different sources of information.

2.2. Learning-based AR visualization

In the present work, we consider the 3 different sources of information provided by a mobile C-arm combined with a RGB-Depth Kinect sensor. The resulting color, depth and X-ray images are represented by their respective intensity functions $\mathbf{I} : \Omega \to \mathbb{R}^3$, $\mathbf{D} : \Omega \to \mathbb{R}$ and $\mathbf{J} : \Omega \to \mathbb{R}$. We assume all images are registered through calibration, so that those functions are defined on the same image domain $\Omega \subset \mathbb{R}^2$. Each pixel $\mathbf{x} \in \Omega$ is associated to three-dimensional value in the CIElab color space, a depth value in mm, and a radiodensity value. Our goal is to create an augmented image $\mathbf{F} : \Omega \to \mathbb{R}^3$ as the fusion of **I** and **J**, taking advantage of the additional depth information contained in **D**. Using a simple method called *alpha blending*, we could construct **F** as a convex combination of both **I** and **J**, ignoring **D**. The same "mixing value" $\alpha \in [0, 1]$ would be then applied to the whole image domain, without taking into account the content of those images. In the present work, we propose to create a pixel-wise alpha mapping based on the semantic content of both images. Ideally, all relevant information needs to be retained and emphasized in the fused image. Based on color, depth and radiodensity information, our novel mixing paradigm can be defined as follows:

$$\mathbf{F}(\mathbf{x}) = \alpha_{\mathbf{I},\mathbf{D},\mathbf{J}}(\mathbf{x})\mathbf{I}(\mathbf{x}) + (1 - \alpha_{\mathbf{I},\mathbf{D},\mathbf{J}}(\mathbf{x}))\mathbf{J}'(\mathbf{x}), \tag{1}$$

where $\mathbf{J}'$ is the function in $\mathbb{R}^3$ that associates a pixel $\mathbf{x}$ to a vector $[\mathbf{J}(\mathbf{x}), \mathbf{J}(\mathbf{x}), \mathbf{J}(\mathbf{x})]^\top$. $\alpha_{\mathbf{I},\mathbf{D}\mathbf{J}}(\mathbf{x})$ is a pixel-wise alpha map that is constructed by taking into account the semantic content, i.e. the relevant objects present in the images to fuse.

2.3. Identifying objects of interest in RGB-D and X-ray images

2.3.1. Related work: object recognition/segmentation in RGB-Depth images

Since the introduction of RGB-Depth sensors such as Kinect, many research have been conducted to tackle the problem of object detection or scene labelling, taking advantage of the combined color and depth information. In the field of pedestrian detection, several works [5–8] propose to combine image intensity, depth and motion cues. In the context of object classification, detection and pose estimation, Sun et al. [9] proposed to detect object from depth and image intensities with a modified Hough transform. More recently, Hinterstoisser et al. introduced in [10] a very fast template matching approach based on so-called multi-modal features extracted from RGB and depth images: they propose to combine color gradient information with surface normals to best describe the templates of the objects of interest. In [11] Silberman et al. also propose to use different type of hybrid features such as RGBD SIFTs within a CRF model in order to segment indoor scenes. In [12], authors tackle the problem of object recognition based on RGB-D images demonstrating that combining color and depth information substantially increase the recognition results. In the following, we will describe in details how we combine color and depth information to identify our objects of interest.

2.3.2. Our approach

Let us first consider both the color **I** and depth images **D** provided by the Kinect sensor, both registered through built-in calibration. Our goal to identify relevant objects within the surgical scene, i.e. objects that belongs to the foreground and to specific classes of interest such as the hands of the surgeon or surgical tools. In this context, we propose to split the task of identifying relevant objects into two subtasks: (1) find candidate foreground objects using the content of the depth image and (2), identify relevant objects using the content of the color image. Formally, each pixel $\mathbf{x}$ needs to be associated with a label $\mathbf{r} \in \left\{0, 1\right\}$, being equal to 1 for relevant objects and 0 otherwise. In our multi-sensor setup, this label $\mathbf{r}$ can be seen as the realization of 2 random variables ($\mathbf{f}$, $\mathbf{c}$), where $\mathbf{f} \in \left\{0, 1\right\}$ represents the observation of a foreground object in the depth image and $\mathbf{c} \in \mathcal{C} = \left\{\textbf{background}, \textbf{surgeon}, \textbf{tool}\right\}$ the observation of classes of interest in the color image. In a probabilistic framework, we aim at modeling the joint distribution $P_{\mathbf{I},\mathbf{D}}(\mathbf{f}, \mathbf{c}|\mathbf{x})$ of a pixel $\mathbf{x}$ to belong to the foreground and to an object class given a depth image **D** and a color image **I**. By decorrelating the observations in depth and color images, we can model this distribution as:

$$P_{\mathbf{I},\mathbf{D}}(\mathbf{f}, \mathbf{c}|\mathbf{x}) = P_{\mathbf{D}}(\mathbf{f}|\mathbf{x})P_{\mathbf{I}}(\mathbf{c}|\mathbf{x}) \tag{2}$$

As modeling the foreground in depth images is ill-posed, we propose to learn instead a background model $P_{\mathbf{D}}(\bar{\mathbf{f}}|\mathbf{x})$ and to use the relation $P_{\mathbf{D}}(\mathbf{f}|\mathbf{x}) = 1 - P_{\mathbf{D}}(\bar{\mathbf{f}}|\mathbf{x})$. Concerning the second term $P_{\mathbf{I}}(\mathbf{c}|\mathbf{x})$, we use a discriminative model based on random forests.