Vision-Based and Marker-Less Surgical Tool Detection and Tracking: a Review of the Literature

David Bouget[a,1,*], Max Allan[b], Danail Stoyanov[b], Pierre Jannin[a]

[a] *Medicis team, INSERM U1099, Université de Rennes 1 LTSI, 35000 Rennes, France*
[b] *Center for Medical Image Computing. University College London, WC1E 6BT London, United Kingdom*

Abstract

In recent years, tremendous progress has been made in surgical practice for example with Minimally Invasive Surgery (MIS). To overcome challenges coming from deported eye-to-hand manipulation, robotic and computer-assisted systems have been developed. Having real-time knowledge of the pose of surgical tools with respect to the surgical camera and underlying anatomy is a key ingredient for such systems. In this paper, we present a review of the literature dealing with vision-based and marker-less surgical tool detection. This paper includes three primary contributions: (1) identification and analysis of data-sets used for developing and testing detection algorithms, (2) in-depth comparison of surgical tool detection methods from the feature extraction process to the model learning strategy and highlight existing shortcomings, and (3) analysis of validation techniques employed to obtain detection performance results and establish comparison between surgical tool detectors. The papers included in the review were selected through PubMed and Google Scholar searches using the keywords: "surgical tool detection", "surgical tool tracking", "surgical instrument detection" and "surgical instrument tracking" limiting results to the year range 2000 - 2015. Our study shows that despite significant progress over the years, the lack of established surgical tool data-sets, and reference format for performance assessment and method ranking is preventing faster improvement.

Keywords: tool detection, object detection, data-set, validation, endoscopic/microscopic images.

1. Introduction

Technological advances have had a considerable impact on modern surgical practice. In particular, the miniaturisation of surgical instruments and advanced instrument design to enable dexterous tissue manipulation have been key drivers behind reducing surgical trauma and giving rise to Minimally Invasive Surgery (MIS) (Dogangil et al., 2010; Davis, 2000; Cleary & Nguyen, 2001). In MIS the surgeon accesses the surgical site through trocar ports, illumination is delivered via optical fibres or light-emitting diodes (LED) and the anatomy is observed through a digital video signal either from a CMOS sensor in the body or an external camera connected to a series of lenses integrated in a laparoscope. By reducing the access incisions and trauma caused by surgery, MIS has led to significant patient benefits and is likely to continue to be one of the most important criteria to the evolution of surgical techniques (Darzi & Mackay, 2002). Specialized surgical instruments are required in MIS to give the surgeon the ability to manipulate the internal anatomy, dissect, ablate and suture tissues. Most recently, such instruments have become robotics manipulators. Mastering the control and use of MIS tools and techniques takes significant training and requires the acquisition of new skills compared to open surgical approaches (Van der Meijden & Schijven, 2009). The MIS instruments deliver a reduced sense of touch from the surgical site, the endoscopic camera restricts the field-of-view (FoV) and localisation (Baumhauer et al., 2008), and the normal hand-motor axis is augmented. As well as impacting the operating surgeon, the introduction of new equipment and devices enabling MIS within the operating theatre means that the whole clinical team must be trained and qualified to operate within the augmented environment in order to avoid preventable adverse events (Kohn et al., 2000). This can have complex implications on clinical training periods and costs, the management of clinical facilities, and ultimately to patient outcomes.

To overcome some of these challenges, computer-assisted intervention (CAI) systems attempt to make effective use of pre-operative and intra-operative patient specific information from different sources, sensors and imaging modalities and to enhance the workflow, ergonomics, control and navigation capabilities during surgery (Mirota et al., 2011; Stoyanov, 2012). A common requirement and difficult practical challenge for CAI systems is to have real-time

[*] Corresponding author
Email addresses: bougetd@gmail.com (David Bouget), maximilian.allan.11@ucl.ac.uk (Max Allan), danail.stoyanov@ucl.ac.uk (Danail Stoyanov), pierre.jannin@univ-rennes1.fr (Pierre Jannin)
[1] Present address: Department Mechanical Engineering, K.U.Leuven, 3001 Heverlee, Belgium.

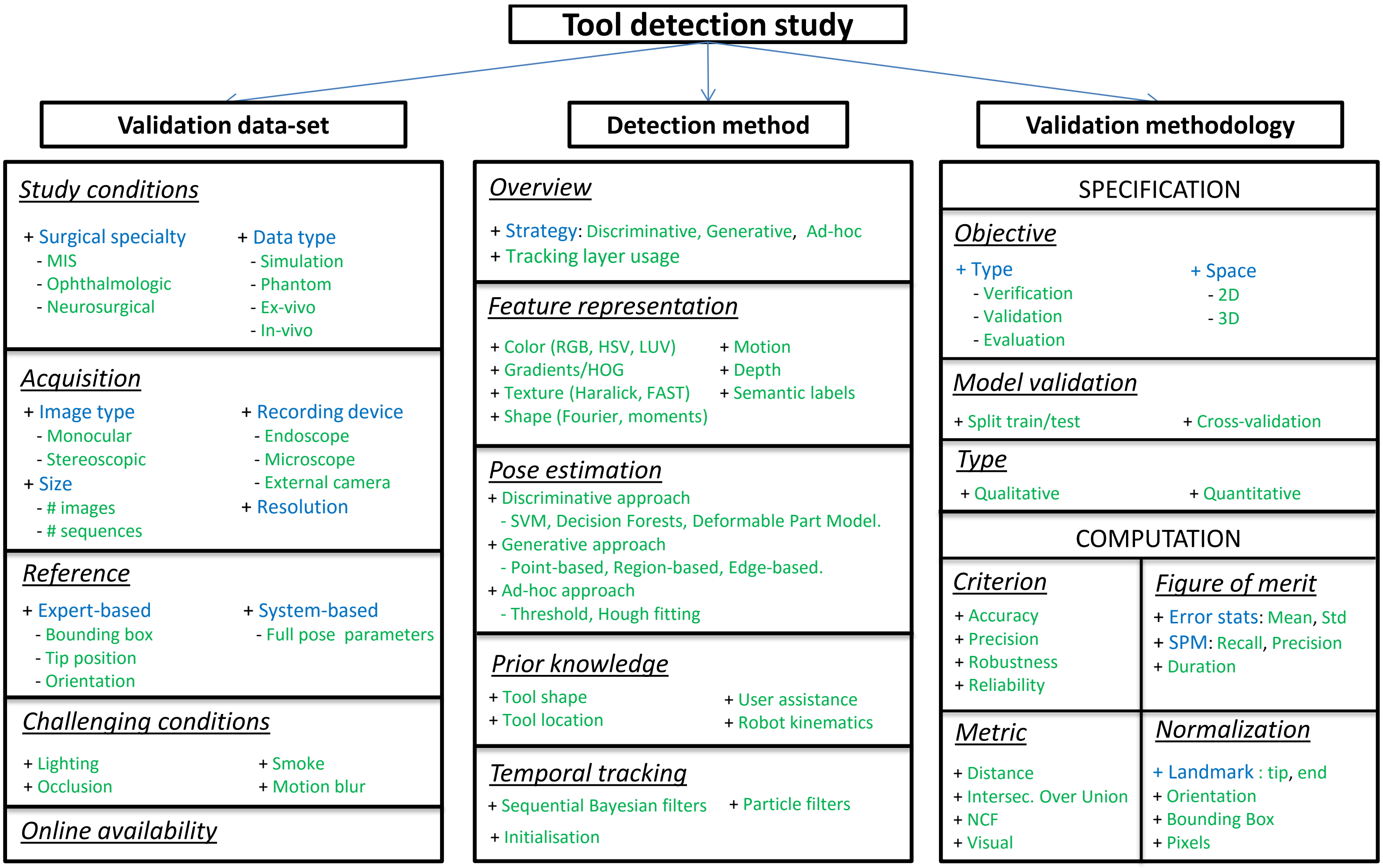


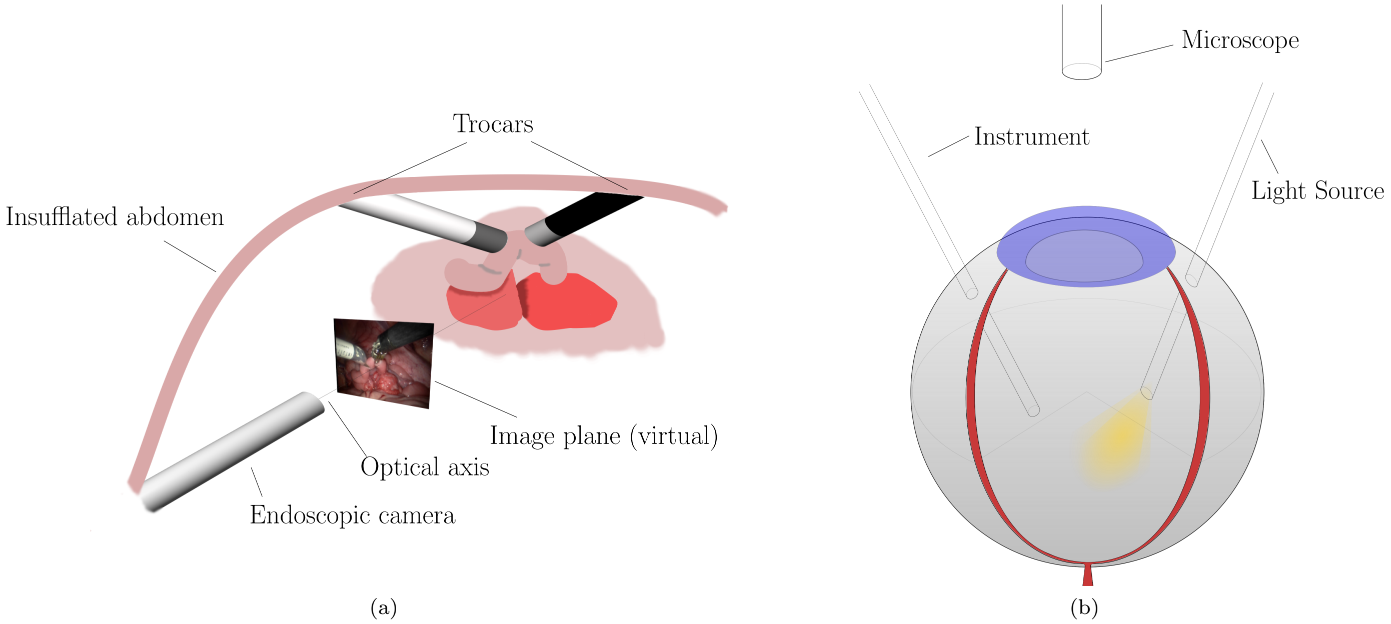
Figure 1: Overview graph describing the strategy followed in this literature review. Each study is classified according to the three main categories: validation data-set, detection methods, and validation methodology. Each category is further sub-divided into multiple components. Sub-division color code: underlined and italic black text represents generic components, blue text corresponds to elements within each component, and green text represents possible values than can be instantiated.

knowledge of the pose of the surgical tools with respect to the anatomy and any imaging information. Different approaches for instrumental localisation have been investigated including electro-magnetic (EM) (Lahanas et al., 2015; Fried et al., 1997) and optical tracking (Elfring et al., 2010), robot kinematics (Reiter et al., 2012b) and image-based tracking in endoscopic images, ultrasound (US) (Hu et al., 2009) and fluoroscopy (Weese et al., 1997). Image-based approaches are highly attractive because they do not require modification to the instrument design or the operating theatre and they can provide positional and motion information directly within the coordinate frame of the images used by the surgeon to operate. A major challenge for image-based techniques is robustness and in particular to the diverse range of surgical specialisations and conditions that may affect image quality and visibility. With this paper we review the current state-of-the-art in image-based and marker-less surgical instrument detection and tracking, focusing on the aspects of prior work. Our major contributions are threefold:

- Summarising the different datasets that are available within the community as well as cohesion and convergence towards a common set of annotations following a standard format;
- Algorithmic review highlighting the various advantages and disadvantages of each method. There are currently no comprehensive reviews on surgical instrument detection which hinders new researchers from learning about the field and additionally prevents cross-pollination of ideas between research groups;
- Analysing the validation methodologies that have been used to produce detection results because currently there is limited consensus on a common reference format for ground truth data or comparison between methods. However, an attempt has recently been made to alleviate this problem with the introduction of the ‘Instrument Detection and Tracking’ challenge at the Endoscopic Vision workshop at MICCAI 2015.

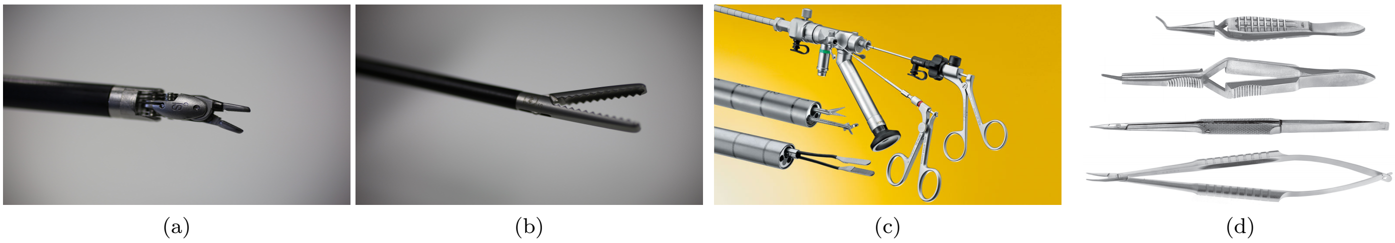
1.1. Review Introduction

For the review, we carried out systematic searches using the Google Scholar and PubMed databases using the keywords: “surgical tool detection”, “surgical tool tracking”, “surgical instrument detection” and “surgical instrument tracking”. In addition to the initial search results, we



(a) (b)

Figure 2: Examples of surgical setups. (a) Minimally invasive surgery setup where surgical instruments enter the cavity through *Trocars*, used to create insertion points as well as facilitate access to the interior thanks to their hollow body. The working cavity is *insufflated* with CO_2 to increase the available space. The operative field is viewed through a camera which is modelled using the usual *pinhole camera* model. (b) Retinal microsurgery setup. The microscope is positioned looking vertically downwards towards the eye with visualisation of the working space provided through the eye's natural opening. Miniaturised instruments and a light source are passed through incisions in the eyeball.



(a) (b) (c) (d)

Figure 3: Examples of surgical tools used in different surgical contexts. (a) da Vinci articulated robotic instrument. (b) Rigid laparoscopic instrument. (c) Neurosurgical instrument. Image modified from ©2015 Karl Stortz GmbH & Co. (d) Retinal microsurgery instrument. Image modified from ©2015 ASICO, LLC.

followed the citations of the obtained papers and all peer reviewed English language publications between 2000 and 2015 were considered. To maintain a reasonable methodological scope, we explicitly focused on papers describing image-based and marker-less surgical tool detection techniques. Other approaches more reliant on the use of external markers, while being not in the scope of this review still represent an important portion of the literature and we provide an overview of such methods in Section 5. We considered methods applied to any surgical field using image data from any type of surgical camera (e.g. endoscope and microscope). A total of twenty-eight publications form the methodological basis for this review and we describe and classify each prior work in three categories: (*i*) validation data-set, (*ii*) detection methods, and (*iii*) validation methodology. The diagram in Fig. 1 shows the subdivision and structure of each category and our systematic methodology.

1.2. Preliminaries

The geometry of imaging a surgical instrument during surgery is shown schematically in Fig. 2a and Fig. 2b for MIS using an endoscope and for retinal microsurgery using a surgical microscope. Endoscopes and surgical microscopes tend to be the most common surgical cameras, and can appear in monocular and stereo variations. The surgical camera is modelled as a pinhole projective camera and its coordinate system is taken as the reference coordinate system. We define detection as the estimation of a set of pose parameters which describe the position and orientation of a surgical instrument in this reference coordinate system. These parameters can for example be (x, y) translations, rotation and scale if working solely in the 2D space of the image plane or alternatively this can extend to (x, y, z) and roll, pitch, yaw of 3D pose parameters. We assume a left handed coordinate system to describe the camera coordinate system with the z axis aligned with the optical axis, unless stated otherwise.

In our review of validation data-sets and methodology components, we refer to terminologies described in (Jannin et al., 2006) and (Jannin & Korb, 2008) where *validation* refers to assessing that a method fulfils the purpose for which it was intended, as opposed to *verification* which assesses that a method is built according to its specifications, and *evaluation* consisting of assessing that the method is accepted by the end-users and is reliable for a specific purpose.

In Fig. 3, surgical tools used in different setups and for different procedures are displayed where two categories emerge. First, instruments are deeply articulated and enable 6 degree of freedom (DOF) movements, such as da Vinci robotic instruments employed for minimally invasive procedures. In the second category, instruments employed can be rigid or articulated with multiple parts, usually for eye-surgery and neurosurgery.

2. Validation Datasets

To describe validation data-sets, we propose to rely on four categories of information: the study conditions in which data have been acquired, the amount of data and its type, the range of challenging visual conditions covered by the data, and the type of data annotation provided. The majority of studies covered in this review focusses solely on its associated data-set, with little cross pollination of datasets between studies. Table 1 provides an overview of validation data-sets and Fig. 4 contains examples from existing data-sets. Amongst them, few are available online and in such cases a link towards the website hosting the data-set is provided. Overall, some information might be missing or inaccurate depending on the level of detail introduced in the corresponding publication and the online availability of the data-set.

2.1. Study Conditions

This component aims to describe the surgical context of each study through characteristics including the assessment scenario, location, environment, and operator (Jannin & Korb, 2008). To cover the identification of the surgical context, we report the surgical specialty as assessment scenario and the data type as assessment environment.

2.1.1. Surgical Specialty

Surgical tool detection has been applied to different surgical specialities, but minimally invasive surgeries including endoscopic and laparoscopic procedures have been the main focus. Surgical examples include hysterectomy (Kumar et al., 2013b), cholecystectomy (McKenna et al., 2005), nephrectomy (Reiter & Allen, 2010) and pelvic (Sznitman et al., 2014). The majority are performed on humans, however in some cases data have been collected on porcine experiments (Pezzementi et al., 2009; Reiter et al., 2012c). Far behind, eye-surgery is the second most studied surgical specialty, especially with retinal micro-surgeries (Burschka et al., 2005; Richa et al., 2011a; Sznitman et al., 2012). Finally, two studies only have investigated tool detection over neurosurgical data, in a spine surgery context (Sznitman et al., 2014) and for brain tumour removal procedures (Bouget et al., 2015).

Outside of the operating room, the scenario does not follow a surgical procedure anymore and is most of the time not specified. In general, specific surgical motions have been targeted (e.g. rapid motion (Richa et al., 2011a)) while generic motions to capture a wide range of poses have also been considered (e.g. (Allan et al., 2013)).

2.1.2. Data Type

The data type component is used to position the data-set along the control versus clinical realism continuum (Jannin & Korb, 2008). The control represents the access to the ground-truth and a gain of control is coupled to a loss of realism, which is represented as a continuum. Four data acquisition types can be identified: simulation, phantom, ex-vivo, and in-vivo. The simulation type represents one end of the continuum with full control and low clinical realism while the in-vivo type represents the other end of the continuum with no control and full clinical realism.

Simulation data are synthetic and obtained from fully controlled environment where tool models and surgical backgrounds can be mixed at will. A simple black tool mask moved on top of a surgical background (Wolf et al., 2011), rendered tool models undergoing translation and articulation against a black background (Pezzementi et al., 2009) or a virtual endoscope moving over a liver with homogeneous texture (Speidel et al., 2013) are representative examples.

Phantom data are a real-world equivalency to simulation data where real surgical instruments are usually moved in front of phantom surgical backgrounds. They can describe very simple setups such as a real tool moving in front of a white background (Allan et al., 2013) or in front of a real surgical background image (Wolf et al., 2011). Phantom backgrounds can be a little bit more complex, such as a half-sphere painted to resemble the retinal surface in an ophthalmic surgical context (Sznitman et al., 2013). Even real phantom organ models can be used to be as close to real clinical conditions as possible (Speidel et al., 2008).

For ex-vivo data, few cases have been reported, with experiments on lamb liver tissue sample (Allan et al., 2013), anatomical structures (Speidel et al., 2013), or cadavers (Voros et al., 2007). An important part of the control versus clinical realism continuum is covered by ex-vivo data. Experiments performed on cadavers will be much closer to the in-vivo category whereas stand-alone tissue sample are closer to the phantom category. Yet, data from ex-vivo experiments are overall simpler because of the absence of physiological motion due to cardiac or respiratory cycles.

In-vivo data are the most represented, nearly used in every study. More details about the range of challenging situations captured in such data are reported in Section 2.4.

Table 1:
Validation data-sets: overview of the data-sets used in the literature to validate tool detection methods. **Surgical specialty**: Minimally Invasive Surgery (M), Eye-surgery (E), Neurosurgery (N). **Recording device**: Endoscope (E), Microscope (M), Consumer camera (C), Time-of-Flight camera (T). **Image type**: Monocular (M), Stereoscopic (S).

	Study conditions					Data acquisition					Data reference				Challenging cond.				Availability
	Surg. spec.	Data type				Rec. device	Image type	Resolution	Size		Manual			System	Lighting	Occlusion	Smoke	Motion blur	
		Simulation	Phantom	Ex-vivo	In-vivo				# Images	# Seq.	Bound. box	Tip pos.	Pose						
(Allan et al., 2013)	M	✔				-	-	720 × 288	-	15				✔					
			✔				M		100	1		✔		✔					
				✔			M		500	1			✔	✔					
					✔		M		97	6	✔								✔
(Allan et al., 2014)	M			✔		E	S	1920 × 1080	400	1			✔	✔					✔
(Allan et al., 2015)	M			✔		E	S	720 × 576	1000	1			✔	✔					✔
(Alsheakhali et al., 2015)	R				✔	M	M	1920 × 1080	400	1		✔	✔						
(Bouget et al., 2015)	N				✔	M	M	612 × 460	2476	14	✔	✔	✔		✔	✔		✔	✔
(Burschka et al., 2005)	E		✔			-	S	-	< 500	1				✔					
	M		✔					-	-	1				✔				✔	
(Cano et al., 2008)	M		✔			E	M	-	550	2				✔					
(Doignon et al., 2005)	M				✔	E	M	640 × 480	52	1					✔				
(Doignon et al., 2007)	M		✔			E	M	-	30	1									
					✔			-	52	1					✔				
(Haase et al., 2013)	M			✔		E	S	640 × 480	30	2		✔							
						T	-	64 × 50											
				✔		E	S	640 × 480	4	4		✔				✔			
						T	-	64 × 50											
(Kumar et al., 2013b)	M				✔	E	M		3k2	16	✔				✔	✔	✔	✔	✔
(Li et al., 2014)	E				✔	M	M	640 × 480	1k5	4		✔			✔			✔	✔
	M					-		-	1k	1	✔					✔			✔
(McKenna et al., 2005)	M				✔	E	M	720 × 576	> 835	2	✔								
(Pezzementi et al., 2009)	M	✔				-	S	640 × 480	> 100	1									
	E		✔			M			> 65	1									
	M				✔	E			> 175	1									
(Reiter & Allen, 2010)	M				✔	-	-	-	> 137	1					✔			✔	
(Reiter et al., 2012c)	M				✔	E	S	-	> 1*k*6	5	✚				✔				
(Reiter et al., 2012a)	M				✔	E	S	-	> 1*k*6	5	✚				✔				
(Richa et al., 2011a)	E		✔			M	S	1*k*6 × 1*k*2	1k1	2					✔	✔		✔	
					✔				-	2									
(Rieke et al., 2015)	E				✔	M	M	1920 × 1080	800	4	✔		✔		✔				
(Speidel et al., 2006)	M				✔	E	M	-	> 60	1	✔								
(Speidel et al., 2008)	M			✔		E	S	768 × 576	560	6	✔	✔			✔				
				✔					100	100		✔				✔			
(Speidel et al., 2014)	M				✔	E	S	320 × 240	542	6		✔				✔	✔		
(Sznitman et al., 2012)	E				✔	M	M	640 × 480	1k5	4		✔			✔			✔	✔
	M					-		-	1k	1	✔					✔			✔
(Sznitman et al., 2013)	E		✔			C	M	1*k*6 × 1*k*2	> 400	1		✔	✔		✔			✔	✔
					✔				850	1		✔	✔		✔		✔	✔	✔
(Sznitman et al., 2014)	M				✔	-	M	640 × 360	1390	2	✔							✔	✔
	N				✔	-		320 × 240	472	1	✔							✔	✔
(Voros et al., 2007)	M				✔	E	M	200 × 100	6	6		✔			✔	✔	✔		
				✔					45	1									
(Wolf et al., 2011)	M	✔				-	M	-	1050	1			✔	✔					
			✔			E		768 × 576	1050	1			✔	✔					
(Zhou & Payandeh, 2014)	M		✔		✔	E	M	-	2000	10					✔	✔	✔	✔	

2.1.3. *Operator and Location*

The remaining study conditions parameters are typically inherent to the data type. For in-vivo data-sets, the operator manipulating surgical tools is the surgeon or sometimes a member of the medical staff and the location is an operating room. In case of ex-vivo or phantom data-sets, the operator is usually a non-medical expert and the location is at best a simulated operating room but most of the times a standard laboratory outside of an hospital.

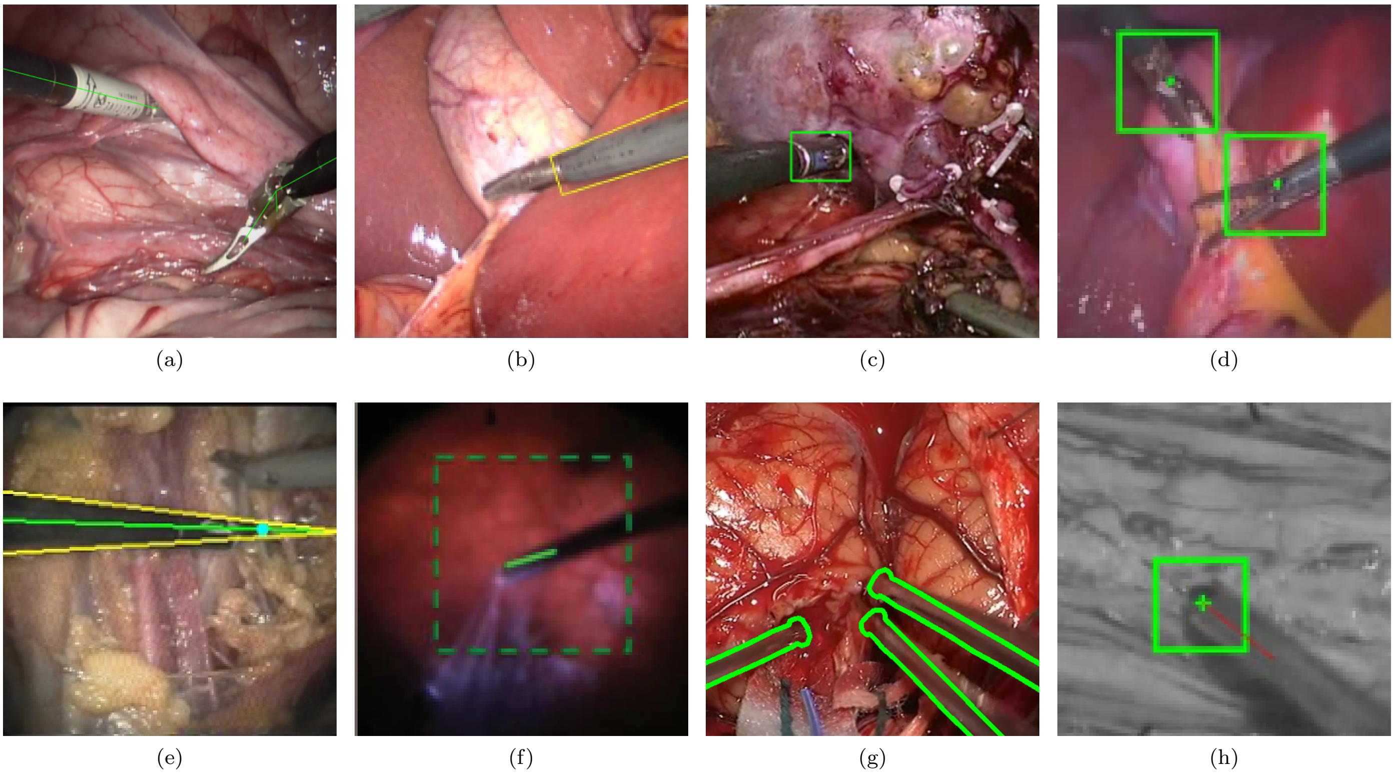
2.2. *Data Acquisition*

The second component aims at describing the data acquisition strategy through the type of recording device employed, the image type and resolution, and the data-set size either as number of images or number of video sequences.

2.2.1. *Video acquisition source*

Video acquisition sources are intrinsically connected to studied surgical specialities as consumer cameras have scarcely been brought into operating rooms. The recording device element is evidently not relevant for simulated data-set where only a computer is needed. For minimally invasive surgery, the most common surgical device in the studied papers is the da Vinci Surgical System [2], equipped with an endoscope as recording device (noted as *E* in the table). In eye-surgery and neuro-surgery, most procedures are performed under a surgical microscope capable of recording videos (noted as *M* in the table).
In spite of constraints in positioning cameras at will due to patient safety concerns, different recording setups have been proposed. Sznitman et al. (Sznitman et al., 2013) managed to couple a consumer camera to a microscope (noted as *C* in the table) and Haase et al. (Haase et al., 2013) used a Time-of-Flight camera (noted as *T* in the table) which they coupled with a 3D endoscope.

[2] ©2012 Intuitive Surgical



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4: Example images from different surgical tool data-sets, with overlaid detections from corresponding papers represented either in green or yellow. (a)-(e) Minimally Invasive Surgery from (Allan et al., 2013), (McKenna et al., 2005), (Reiter & Allen, 2010), (Sznitman et al., 2013) and (Voros et al., 2007) respectively. (f) Eye surgery (Sznitman et al., 2012). (g)-(h) Neurosurgery from (Bouget et al., 2015) and (Sznitman et al., 2014) respectively.

2.2.2. Image Type

Depending on hardware capabilities of video acquisition sources, two image types are available: monocular and stereoscopic (noted as M and S respectively in the table).

Monocular represents single images, enabling to retrieve 2D positions in the image referential only and is the most represented image type such as in (Doignon et al., 2007; Sznitman et al., 2013; Wolf et al., 2011).

Stereoscopic represents a pair of monocular images enabling the retrieval of depth estimates using epipolar geometry. They have been for example used in (Burschka et al., 2005; Pezzementi et al., 2009; Reiter et al., 2012c).

2.2.3. Image Resolution

A high variability can be noted in reported image resolutions, ranging from low resolution (e.g. 200×100 pixels (Voros et al., 2007)) to high resolution (e.g. 1920×1080 pixels (Allan et al., 2014)) images.

2.2.4. Data-set Size

The amount of data, represented as a number of images, can be expressed in different levels of magnitude: small, medium and large, with respect to the largest data-set used in the literature proposed by Kumar et al. (Kumar et al., 2013b). Small data-sets contain less than a hundred images (Doignon et al., 2007; Haase et al., 2013; Speidel et al., 2006), while medium data-sets range from a hundred to a thousand images (Allan et al., 2013; McKenna et al., 2005; Sznitman et al., 2013). Finally, large data-sets incorporate more than a thousand images (Bouget et al., 2015; Kumar et al., 2013b; Reiter et al., 2012c).

The variability within a data-set is also a paramount factor and is represented by the number of video sequences from which images originated. A sequence is a video recorded on the same subject and with the same imaging modality where only zoom, orientation, and illumination parameters can vary. Most of the studies, especially early works in the field, present a data-set made of one video sequence only (Burschka et al., 2005; Pezzementi et al., 2009; Reiter & Allen, 2010). However, recent works such as (Allan et al., 2013; Bouget et al., 2015; Kumar et al., 2013b) are proposing more than 5-6 sequences thus introducing greater diversity in the data pool.

2.2.5. Tools Statistics

In addition to the number of different video sequences, a second level of variability involves the number of different surgical tools and their occurrences in the data-set. Those information being only scarcely accessible, they are not displayed in the table. However, we propose to report here what we managed to gather.

Regarding surgical tools diversity, studies were mostly focusing on tubular-shaped tools such as a forceps (Pezzementi et al., 2009), a large needle driver (Reiter et al., 2012c), a cylindrical needle-holder (Doignon et al., 2004), or standard articulated tools from da Vinci (Kumar et al., 2013b). Many times, only the generic term of "endoscopic tools" or "tools" is mentioned (Haase et al., 2013).

Some data-sets only feature one surgical instrument, especially in phantom conditions (Doignon et al., 2007; Wolf et al., 2011). Two simultaneous tools are widely featured, especially in a context of MIS (McKenna et al., 2005). More than two surgical tools is very unlikely, mostly because of the nature of minimally invasive surgeries performed using a da Vinci and displaying a maximum of two tools at the same time. Speidel et al. (Speidel et al., 2014) proposed a data-set with up to three tools simultaneously visible for a total of four different tools while Bouget et al. (Bouget et al., 2015) also introduced a data-set showcasing more than two tools at the same time for up to seven different tools in total.

Unfortunately, tool occurrences, overlapping occurrences, orientation, number of simultaneous tool distributions, or any kind of extended statistics, are not available in any paper with the exception of (Bouget et al., 2015).

2.3. Data Reference Creation

The final objective being to estimate surgical tool pose in images, having a reference for said positions, assumed to be close to the correct result (Jannin et al., 2006), is necessary. They can be obtained in two ways: either manually or automatically. The manual approach being widely employed as not requiring the installation of additional external sensors.

The favored approach to obtain automatic annotations is to use an Optotrack optical localizer (Allan et al., 2013; Burschka et al., 2005; Wolf et al., 2011). For simulated data-sets, tool pose parameters are inherently known by the computer setting up the simulation (Wolf et al., 2011; Speidel et al., 2013).

Regarding manual annotations, most of the time tool-tip positions are involved such as in (Haase et al., 2013; Speidel et al., 2014; Wolf et al., 2011), along with bounding boxes around surgical tools (Kumar et al., 2013b; Speidel et al., 2006; Sznitman et al., 2014) or parts of surgical tools (Reiter et al., 2012a) (represented in the table by ✚). Occasionally, variants of bounding boxes are used such as bounding polygons (Bouget et al., 2015) or pixel-wise labelling (McKenna et al., 2005). Extended pose parameters can also be documented such as tool orientation (Bouget et al., 2015), length, or entry point in the image (Sznitman et al., 2013).

2.4. Challenging Conditions

The third and last level of data variability corresponds to the range of challenging conditions captured. Data-sets may cover a wide range of appearance and lighting scenarios (Reiter & Allen, 2010; Reiter et al., 2012c) sometimes inducing shadows (Sznitman et al., 2013), include occlusions (Speidel et al., 2014), rapid appearance changes (Reiter & Allen, 2010), smoke (Sznitman et al., 2013), specular reflections (Kumar et al., 2013b), blur (Sznitman et al., 2012) or blood spatter (Haase et al., 2013). While sometimes data-sets explicitly do not cover any challenging situations (McKenna et al., 2005).

2.5. Online Availability

Having presented the composition and shortcomings of previously used surgical tool data-sets, a major issue remains to be addressed: their availability. For efficient methodological bench-marking and comparison, the ability to test on multiple datasets within a unified framework is paramount. In the following is a list of web-sites where data-sets can be freely downloaded.

Data-sets presented by Allan et al. in (Allan et al., 2013, 2014), Sznitman et al. in (Sznitman et al., 2011, 2012, 2013, 2014), Kumar et al. (Kumar et al., 2013b), and finally Bouget et al. (Bouget et al., 2015) are respectively available at[3] [4] [5] [6]. All the remaining data-sets used in previous studies have not been made openly available online. In addition, some data-sets designed for surgical tool detection and tracking have been created but never yet used in a proper study (e.g. (Maier-Hein et al., 2014)).

3. Tool Detection Methods

Detection of any object can be described quite generally as a parameter estimation problem over a set of image features. Broadly there are three strategies which have been used to solve the problem. The first two fit within a more holistic modelling paradigm and are separated into discriminative methods using discrete classification and generative methods which aim to regress the desired parameters in a continuous space. The third strategy encompasses ad-hoc methods that rely on empirical combinations of simple models for detection.

In this section, we review how the image data is condensed into a manageable, low dimensional representation in the form of features and then contrast the three broad approaches to detection of surgical instruments. After which we introduce prior knowledge and detail how to temporally link detection results together via tracking strategies. To finish, we describe general optimization strategies which can be employed to constrain the detection search space as to obtain faster and/or more accurate detection results.

[3] http://www.surgicalvision.cs.ucl.ac.uk/benchmarking

[4] https://sites.google.com/site/sznitr/code-and-datasets

[5] http://mechatronics.eng.buffalo.edu/research/rMIS_SkillAssessment/PoTE_DataSet.html

[6] http://dbouget.bitbucket.org/2015_tmi_surgical_tool_detection

Table 2:
Tool detection methods: overview of the methods used in the literature to detect surgical tools. **Strategy**: Discriminative (D), Generative (G), Ad-hoc (A).

	Overall		Features								Pose estim.			Prior knowledge				Tracking		
	Strategy	Tracking	Color	Gradients	HOG	Texture	Shape	Motion	Depth	Semantic Labels	Discriminative	Generative	Ad-hoc	Tool shape	Tool location	User assist.	Kinematics	Bayes.	Particle	Initialisation
(Allan et al., 2013)	G		✔				✔			✔		✔		✔	✔					
(Allan et al., 2014)	G	✔	✔			✔	✔			✔		✔		✔	✔	✔		✔		
(Allan et al., 2015)	G	✔	✔				✔	✔		✔		✔		✔	✔	✔		✔		
(Alsheakhali et al., 2015)	A		✔	✔			✔						✔							
(Bouget et al., 2015)	D		✔		✔	✔				✔	✔									
(Burschka et al., 2005)	A D	 ✔	✔ ✔	✔						✔	 ✔		✔	✔			 ✔	 ✔		
(Cano et al., 2008)	A	✔	✔	✔									✔	✔	✔					✔
(Doignon et al., 2005)	A		✔	✔			✔						✔	✔	✔					
(Doignon et al., 2007)	G			✔								✔		✔	✔					
(Haase et al., 2013)	A		✔	✔					✔				✔	✔	✔					
(Kumar et al., 2013b)	D	✔			✔	✔		✔			✔									✔
(Li et al., 2014)	D	✔		✔		✔					✔					✔				✚
(McKenna et al., 2005)	G	✔	✔							✔		✔		✔	✔				✔	
(Pezzementi et al., 2009)	G		✔			✔						✔		✔						
(Reiter & Allen, 2010)	G	✔	✔			✔						✔		✔		✔				✔
(Reiter et al., 2012c)	G	✔	✔			✔						✔					✔	✔		
(Reiter et al., 2012a)	D		✔	✔		✔				✔	✔			✔	✔		✔			
(Richa et al., 2011a)	D	✔	✔								✔									✔
(Rieke et al., 2015)	D	✔	✔		✔						✔				✔					✔
(Speidel et al., 2006)	G	✔	✔									✔			✔				✔	
(Speidel et al., 2008)	A		✔						✔	✔			✔	✔						
(Speidel et al., 2014)	A	✔	✔					✔	✔				✔		✔					✔
(Sznitman et al., 2012)	D	✔		✔							✔					✔				✚
(Sznitman et al., 2013)	G D	✔	✔ ✔				✔ ✔				 ✔	✔						✔		
(Sznitman et al., 2014)	A			✔						✔			✔	✔						
(Voros et al., 2007)	A			✔			✔						✔	✔		✔				
(Wolf et al., 2011)	G	✔		✔			✔					✔		✔		✔			✔	
(Zhou & Payandeh, 2014)	A	✔		✔									✔	✔				✔		

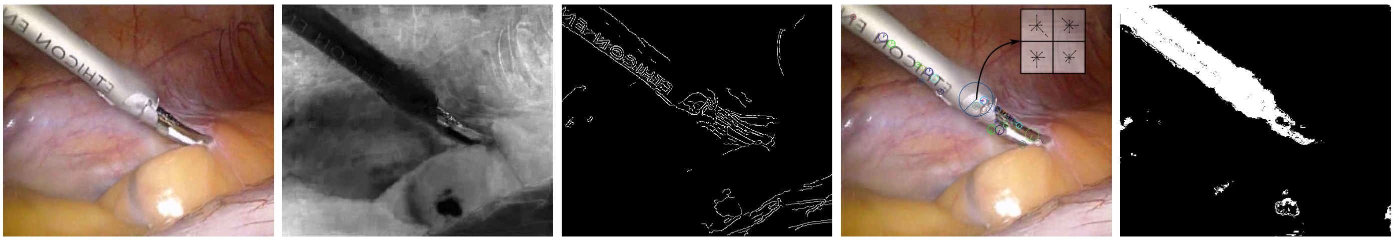
3.1. Feature Representation

The results of linear or non-linear transformations of the input image are called *features*. Features computed over the input image and aggregated into specific representations serve as a basis for object-specific model learning and classification. The selection of sufficiently distinguishable natural features is a challenging aspect of a detection system. Combinations of features provide a potentially more discriminative feature space but require more computational power and increase size of the required training set. Dimensionality reduction techniques can be employed to enforce cheap enough feature representation computation while still providing the accuracy required to avoid data association errors. Only in (Pezzementi et al., 2009), authors relied on one of the main existing strategies; the Linear Discriminant Analysis (LDA) (McLachlan, 2004). The second set of columns in Table 2 indicates feature types used in each study. Features are reported as general category of image content extracted and not as particular instances.

3.1.1. Color

The most popular and widespread of the natural features is color. Nearly all of the existing methods for detecting surgical instruments in images use color information as the primary or sole visual aid due to its ease of computation and simplicity (see Fig. 5b). Compared to the color-based fiducials mentioned in Section 5, employing natural color features is much more challenging due to visual ambiguities created by shadows and lighting.

The RGB colorspace was initially investigated for tool detection as part of the framework developed by Lee et al. (Lee et al., 1994) in a MIS surgical context. More recently, it has been directly used in (McKenna et al., 2005; Reiter & Allen, 2010; Sznitman et al., 2013; Zhou & Payandeh, 2014). Yet, RGB has often been supplanted by the HSV/HSL colorspace (Doignon et al., 2004; Speidel et al., 2009), offering a separation between the chromaticity and the luminance component. By decoupling luminosity from other components, more robustness is provided towards lighting changes. The CIE L*A*B* space which is closely modelled on human visual perception has also been used (Alsheakhali et al., 2015). This space allows a wider range of possible colors than RGB, but come with the tradeoff that more than 8 bits are needed to each channel, increasing processing time. To quantify which colorspaces provide the best discrimination between surgical instruments and tissue, a colorspace analysis can be performed (Allan et al., 2013). Using the metrics of random forest variable importance (Verikas et al., 2011) and Bhattacharyya Distance (Bhattachayya, 1943), they evaluated each channel of RGB, HSV, Cie XYZ as well as Opponent 1 and Opponent 2. Hue and Saturation along with Opponent 1 and 2 were found to provide the best discriminative power. One additional challenge that occurs when comparison within non-Euclidean colorspaces such as HSV is that the common Euclidean distance metric is not valid. To address this, the *coneHSV* colorspace has been used in (Pezzementi et al., 2009), expressing the HSV colorspace transformed as $(V, S\cos(H), S\sin(H))$. An alternative to leveraging standard colorspaces is possible through the



(a) (b) (c) (d) (e)

Figure 5: Examples of image features. (a) shows an image from a typical minimally invasive procedure captured through a laparoscope. (b) shows the frame transformed into the saturation color space, which is often effective at highlighting metallic objects, (c) shows edge features, (d) shows extracted texture features (SIFT), and (e) shows a semantic labelling map.

concept of Color Names (CN) (Van De Weijer et al., 2009). Pixel values, expressed in the L*A*B* colorspace are used within a model-based strategy to learn specificities of labelled colors. Such color features have been employed for surgical tool detection in (Bouget et al., 2015).
While being relatively simple to compute, color features have significant shortcomings. Despite the obvious dissimilarity between the red hues of tissue and the monochromacity of instruments, the lighting used in medical environments combined with the smooth tissue surface causes large specular reflections disrupting the white and grey appearance of metallic instruments. This leads to particular challenges when classifying the instruments using color alone.

3.1.2. Gradient

The second most popular feature type revolves around the use of gradients (represented by Gradients and HOG columns in the table). Typically, gradients are generated from color image for example from intensity values or specific colorspace component (e.g. Saturation). Yet, depth maps have also proven to be effective for gradients computation (Haase et al., 2013), and potentially other image types can be used as input.

Gradient features can be extracted either through the computation of image derivatives in x- and y- axis (Wolf et al., 2011) or by performing Sobel filtering (Haase et al., 2013). However, directly leveraging such low-level edge information is somewhat difficult, hence most of the time it has served as input for the Hough transform in order to efficiently retrieve lines in the image (Haase et al., 2013; Voros et al., 2007; Zhou & Payandeh, 2014).

A more robust representation of gradients exists in the form of the well known Histograms of Oriented Gradients (HOG) (Dalal & Triggs, 2005). Typically, not all the orientations, or oriented gradients, are represented but rather a discrete number corresponding to the amount of bins of the histogram. Nevertheless, fairly few tool detection approaches have been relying on this representation (Bouget et al., 2015; Kumar et al., 2013b; Rieke et al., 2015). The classic number of orientation bins, used as well for other computer vision object detection instances such as pedestrian (Dollár et al., 2009), is six. Additionally, the gradient magnitude is typically added as a complementary seventh channel.

Variants of the HOG framework have been preferred in other studies through the use of edges and dominant orientations (Reiter et al., 2012a; Sznitman et al., 2012, 2014).
In general, those feature representations are useful for describing oriented edges and corners but suffer heavily from noise which is common in medical images. Fig. 5c shows an example set of gradient images which illustrates the level of noise that occurs when trying to extract useful edges.

3.1.3. Texture

More robust representations of gradient features can be achieved by extracting *texture* information which can be defined as periodically repeated local patterns in an image. Originally, texture features have been extracted using filter responses for example through Gabor filtering, via textons (Malik et al., 2001), or Local Binary Patterns (LBP) (Ojala et al., 2002). A popular strategy for object detection lies in *Interest points* detection since the emergence of the highly successful SIFT features (Lowe, 1999) (see Fig. 5d), which has spawned numerous other attempts (Dalal & Triggs, 2005; Bay et al., 2008; Calonder et al., 2010; Ambai & Yoshida, 2011). All are based on the principle that creating histograms of gradient orientation around a particular *keypoint* allows it to be correctly matched when viewed from a different viewpoint. Despite the popularity of these methods in other areas of computer vision, they have not been used extensively in the task of surgical instrument detection. One particularly successful attempt has been made by Reiter et al. in (Reiter et al., 2012c,a), making use of SIFT features learned around the head of da Vinci robotic instruments. Previously, the same authors also tried to make use of FAST corners (Reiter & Allen, 2010). While a similar *keypoint* extraction strategy can be employed, the texture representation diversity can come from the *keypoint* description process. Generally, each *keypoint* extraction strategy is affiliated to its corresponding *keypoint* description process (e.g. SIFT, SURF). However, other description strategies, such as Color-SIFT (Abdel-Hakim et al., 2006) have been assessed by Allan et al. (Allan et al., 2013) for surgical tool detection. Being built into a similar structure to the one of SIFT, the Color-SIFT description strategy provides both color and geometrical invariance.

In Pezzementi et al. (Pezzementi et al., 2009), authors proposed to measure the texture of the image using co-occurrence matrices and to perform the representation using a sub-set of Haralick coefficients. They used four Haralick texture features (Haralick, 1979) based on contrast, correlation, energy and homogeneity of 3×3 image patches. The Haralick features are computed from the gray-level co-occurrence matrix of which 14 difference statistics can be computed. The adjacency criteria for the co-occurrence matrix is modified to consider the average of horizontal, vertical, and diagonally adjacent pixels which provides the features with some informal invariance to rotation. This strategy has been later on integrated into the work of Reiter et al. (Reiter et al., 2012a).

3.1.4. Shape

Amongst the least represented categories, surgical tool detectors can utilize *shape* features, generally represented as a set of numbers produced to describe a given shape. Different approach types can be followed such as region-based, space-domain, and transform-domain shape features (Yang et al., 2008).

Region moments, for instance Hu invariant moments, are very popular amongst region-based shape features. In (Voros et al., 2007), authors relied on the Otsu's thresholding technique (Otsu, 1975) to identify the tool-tip location by finding the optimal separation between instrument and background pixels by computing zeroth-, first-, and second-order cumulative moments. Region moments are mathematically formulated to offer invariance under translation, scale, and rotation for an average computational complexity. However, they provide a very limited robustness towards noise, occlusion or non-rigid deformation, for a highly redundant information extracted.

Wavelet transform features, such as the Haar wavelets (Papageorgiou et al., 1998), have been extensively used since the well known Viola-Jones face detector (Viola & Jones, 2004) and employed by Sznitman et al. in their surgical tool detector (Sznitman et al., 2013) and additionally to filter edges using a blob detector in retinal instrument detection Alsheakhali et al. (2015). They give strong responses to sharp directional intensity edges in images by summing the pixel intensities within the filter boundaries. However, they are not as robust to noise as other filters that they approximate (e.g. Gabor filters) and are not as flexible as steerable features which are not restricted to horizontal and vertical gradients.

Within transform-domain shape features, Fourier descriptors have been previously used in (Doignon et al., 2005) to enable better classification of regions as instrument or background. As their color based segmentation method produces several outliers, they are forced to incorporate the shape of the region as part of their evaluation. Fourier descriptors describe the boundary of a region by computing a Fourier component for each pixel in the boundary. Following from the properties of the Fourier transform, this descriptor can be shown to be invariant to rotation, translation, scaling and origin. By extracting the outer contour of a region detected by a color classifier and taking the Euclidean distance between the region's Fourier descriptors, the most similar shape in the image is taken as the one with the minimal distance. The authors also combine the Fourier descriptors with affine invariant region moments (Flusser & Suk, 1993) to improve the robustness of their region detection, again using the Euclidean distance between the moments.

3.1.5. Additional Categories

Less traditional or easy to obtain features have also been scarcely investigated in the literature (noted as *depth* and *motion* in the table). In the recent work from Speidel et al. (Speidel et al., 2014), the authors added motion and disparity as features to better segment instruments from the background. Disparity features enable the use of depth information in the field-of-view and build upon the fact that instruments are typically closer to the camera than tissue surfaces. This feature is of course extensively reliant on the quality of the reconstruction algorithm, typically using color or gradients to match correspondences, but can be extremely useful due to the high-level smoothness constraints they contain. This feature type has also been used in (Haase et al., 2013; Speidel et al., 2008). Another important cue for human perception, which has almost never been incorporated successfully, is *motion*. Motion cues provide a strong discriminative power due to the distinctive motion patterns displayed by the surgical tools. In (Speidel et al., 2014), motion cues were incorporated by taking difference images from deinterlaced images and disparity maps at subsequent timesteps. In (Allan et al., 2015) optical flow features were used to track features on robotic instruments in complement with color and shape features. Lastly, one attempt only has been made towards the use of spatio-temporal features in the work of Kumar et al. (Kumar et al., 2013b). The authors proposed to compute dense optical flow, but for use within a tracking framework.

3.1.6. Semantic Labelling

While all the aforementioned feature types are directly extracted from input images, a higher level of features can be generated in the form of semantic labelling maps by making use of classification algorithms (noted as *semantic labels* in the table, see Fig. 5e). These maps can be generated by learning a relationship between any number of the low level pixel features already addressed and desired object classes. Typically either a simple binary case of an instrument and tissue background separation is modelled (Bouget et al., 2015; McKenna et al., 2005). However, more complex labellings breaking the instrument down into several sub-classes have also been carried out (Allan et al., 2015; Reiter et al., 2012c).

Normally, the simplifying assumption that all pixels are independent serves as basis for semantic classes modelling. These classifications have either been carried out by maximising the conditional distribution over the pixels (Pezzementi et al., 2009; Reiter et al., 2012a; Speidel et al., 2014), or by learning one cascade classifier per class (Bouget et al., 2015; Sznitman et al., 2014). However, more sophisticated methods which model the dependence between neighbouring pixels (Krähenbühl & Koltun, 2012) or include shape and edge based information into the borders between neighbouring classes (Shotton et al., 2009) have been proposed in general computer vision papers. Traditionally, semantic labelling maps have served as direct input to the pose estimation but have never been associated with low-level features in a combined representation. Yet, they can also be used subsequently to the pose estimation as part of a refinement step aiming to remove outliers (Reiter et al., 2012a).

3.2. Pose Estimation

Given a set of extracted features, the central part of detection is the estimation of the parameters which describe the instrument's pose in the image. As illustrated in the third set of columns in Table 2, three main paradigms have been explored and are introduced in detail in the following.

3.2.1. Discriminative Detection

First, joint surgical tool detection and pose estimation has been solved by performing a fully exhaustive search over the input image within a specified range of pose parameters. An inherent supervision is required to generate data-driven surgical tool models, as long as quality training samples.

Methodological Overview. Initially, a model representing the tool of interest to be detected is generated either through a data-driven process or via hand-crafting. Data-driven processes are heavily influenced by computer vision and machine learning lines of work, mostly focusing on object detection instances such as faces or pedestrian (Dollár et al., 2011). The main existing paradigms are HOG+SVM (Dalal & Triggs, 2005), decision forests (Viola & Jones, 2004), Deformable Part Modelling (DPM) (Felzenszwalb et al., 2008) and Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012). A tool model is then traditionally processed over the full input image in a sliding window fashion and for different pose configurations. Parameters searched are typically x and y translation where the bounds on the search space are naturally provided by the image dimensions. To improve the search range completeness, multi-scale and multi-orientation scanning can be additionally performed (Bouget et al., 2015). For each set of parameters, a detection score is returned, indicating the confidence for the tool to be present in a given location of the image and under a specific pose configuration. Eventually, a Non-Maximum Suppression procedure is often performed to limit the number of candidate detections to the most promising ones only (Dollár et al., 2011).

Literature Approaches. Discriminative approaches are very similar in their sliding window implementation whereas their main point of divergence comes from the tool model learning strategy employed.

Decision forests (e.g. random forests), offering the possibility to be used on top of any kind of image feature channels, have been used multiple times. In (Bouget et al., 2015), a Random Forest model is learned over 10 different feature channels, compensating in translation, scale, and orientation. At run-time, an exhaustive exploration of these parameters when evaluating the model is required, potentially leading to a too slow process. Conversely in (Sznitman et al., 2012), authors proposed to rely on a deformable detector (Ali et al., 2012) designed to overcome modelling difficulties regarding deformations and rotations. The designed set of pose-estimator features produces a deformable detector which can learn deformations and rotations from training data, hence not requiring an exhaustive evaluation at run-time. Recently, (Rieke et al., 2015) used Random Forests to track retinal instruments by modelling a template deformation online to estimate the 2D translation between subsequent frames.

The DPM paradigm has been employed once to model a surgical instrument through a latent SVM used in combination with HOG features (Kumar et al., 2013b). The modelling is focusing on the tool-tip often in contact with the tissue and most likely to be visible at all time when the tool is being used. Furthermore, to capture tool articulations, they used a star-structure pictorial model linking root of an object to its parts using deformable springs. Finally, a more standard approach with a linear SVM has been used to learn tool shape models in (Bouget et al., 2015). The modelling focuses on the global shape of the object only and integrates a 2D spatial regularization constraint to enforce shape consistency in the model.

Linear SVM and decision forests detectors present multiple advantages through their theoretical guarantees, good performance, and high-speed capabilities. Being highly extensible and generalizable, they enable to perform tool detection without making any restrictive assumptions about the type, shape, color, or view-point. In addition, decision forests are of particular interest as they automatically perform feature selection which is more robust that hand selected features (Benenson et al., 2013). Yet, the feature representation can incorporate limited assumptions within the training process, inherent to their computation (e.g. U-SURF). Having a pool of features both large and diverse enough almost nullifies such impact, as when employing Convolutional Neural Networks. Even though such frameworks present extensibility and generalization advantages, finding a reasonable amount of training data is necessary. By taking into consideration training pools described in previous works, around 500-1000 positive samples seems

enough to provide with excellent detection results. Evidently, the diversity in positive samples, when multiple video sequences are used, should be taken into account.
While discriminative approaches are built to perform fully exhaustive searches, some authors proposed to only look for the surgical tool in a sub-window of the initial image or for a sub-range of pose parameters. In (Sznitman et al., 2013), an original approach is followed with Active Testing. While it can be seen as a generative approach because of the iterative stochastic optimization, the template matching oracle question can be viewed as a discriminative step. In such case, the output from the generative counter-part is used as prior knowledge to limit the search range. A Sum of Square-Difference (SSD) is performed between the hypothesized pose and the projection of a tool-like color model. The model, using the hypothesized pose, is built with a 1 value at instrument location and 0 elsewhere. The template matching strategy is mimicking the one used in a previous surgical tool detection work by Burschka et al. (Burschka et al., 2005).
In (Reiter et al., 2012a), the Line-Mod template matching strategy is used (Hinterstoisser et al., 2011), relying on an HOG-like object model consisting in discretized orientations from image gradients. To perform model matching, the number of similar orientations between the model and the current image location is estimated as an energy function. Its robustness is based on dominant gradient orientations enabling to perform pose estimation with a limited set of templates. Reiter et al. (Reiter et al., 2012a) made use of a pre-defined 3D CAD model perfectly describing tool joints and derived a set of 2D templates for different articulation configurations. At run-time, not all templates are evaluated but rather a sub-set selected using the robot kinematics defining a reduced range of articulation configurations. In (Richa et al., 2011a), a brute-force template matching is used with Mutual Information as model template. The matching is performed using a measure based on weighted mutual information.
Most of the time, only the tip region has been modelled, but no-one tried to evaluate the impact of modelling tools with different lengths. In addition, multiple tool articulation configurations are not easy to process in real-time with discriminative approaches as the computation would quickly get too huge. Only a limited number of pose parameters can be obtained with such “brute-force” approaches.

3.2.2. Generative Detection

The alternative approach to detection is to handle the parameter estimation problem with generative models. These methods are typically regression based and perform iterative local searches for a solution.

Methodological Overview. The process involves constructing a model f which enables the chosen feature representation of the image to be *predicted* given an estimate of the model parameters. This can either be described in terms of probabilities where a joint distribution over the model parameters and data is learned from training examples or alternatively a sampling procedure is used to generate feature representations from the model which are then compared to the data with a loss function.

Point-Based Approaches. A popular generative method for detecting the pose of a target object is to minimize a distance metric between a model of salient features on the target object’s surface and detections of these features in images or video. This problem is commonly referred to in the computer vision literature as the *PnP problem* and its solutions have been extensively studied (Lepetit et al., 2009; Hartley & Zisserman, 2004; Moreno-Noguer et al., 2007). The generative model for this type of method is typically quite simple and involves a pinhole camera projection of each 3D surface feature to the image plane. The instrument pose parameters are then sampled to minimise an error metric between the projections and matched correspondences in the image.

Using feature points to detect instruments first involves choosing the features with known coordinates on the model surface which are going to be matched to detected features in a given image. Normally these features are based on histograms of local gradient orientations (Allan et al., 2014) or combinations of color and gradient features (Reiter et al., 2012d). Building the geometric and appearance models for the 3D surface points can either be done offline, where the locations and appearance of the salient points are learned for a geometric model of the object (Reiter et al., 2012d) or online where frame-to-frame tracking is achieved by matching points between images (Allan et al., 2014).

Model/Region-Based Approaches. Region-based detection provides a much larger set of constraints than point based methods as they use the full visible appearance of the instrument. Formulating the generative model for this task involves constructing a surface appearance model and a shape model. The shape model constrains the contours which divide the image up into regions and the surface appearance model is used to assess whether a proposed region agrees with the underlying image data. Surface appearance is normally modelled with a bag-of-pixels type model where spatial relationships between features on the surface are discarded. Pezzementi et al. (Pezzementi et al., 2009) developed one of the earliest method in this area by learning gradient and color based features which were used to produce a probabilistic region image. Color features have also been previously used in (Allan et al., 2013, 2014, 2015). The most common method of constraining the shape of regions is with 3D models (Allan et al., 2013, 2014; Pezzementi et al., 2009) where projections are used to generate segmenting contours. These strict prior models make the optimization more complex and computationally expensive than working with flexible 2D contours (Bibby & Reid, 2008) but allow the estimation of 3D pose without a complex step of back projecting the contour to find the corresponding pose (Cashman & Fitzgibbon, 2013a) and additionally allow the exact shape of the contour to be build into the optimization. Simpler models have also been used in (McKenna et al., 2005) based on cylinders, which work effectively for simpler instruments such as laparoscopic models but do not scale well when considering more complex robotic instruments. 2D models have also been used which provide much simpler optimization frameworks at the cost of lower modelling accuracy. 2D bounding boxes have been used as a shape model in (Lee et al., 1994; Lo et al., 2003).

Edge-Based Approaches. Edge-based methods make use of directional ensembles of gradient features to detect the pose of objects. Typically the generative model uses a 3D model of the object and a rendering pipeline to predict the location of edges in the image, given a current pose estimation. If only the external edge of the object are used, this method bears some similarity to the region based methods but additionally internal edges can be used to better constrain the estimate. These edges are matched to strong gradients in the image either using an explicit 1D correspondence search between the model edge and the closest edges in the image or alternatively implicit correspondences can be applied with signed distance function representations (See Fig. 6).

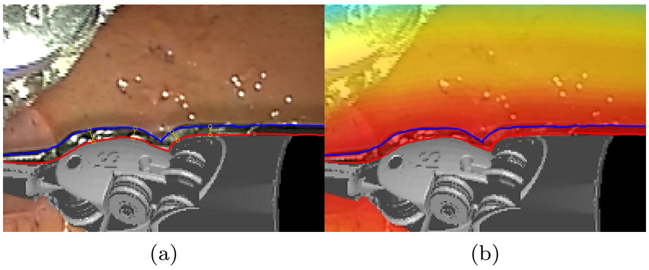


Figure 6: (a) shows explicit correspondences in yellow between the edge in the image (blue) and the edge from the model (red). (b) shows an alternative situation where the corresponces are implicitly defined by a distance function which is represented with a rainbow color map where red to blue indicates distances that are further from the edge.

One of the significant challenges in edge based methods is the difficulty in finding valid correspondences between the edges of the model and the edges found in the images, as typical medical images contain many spurious edges. A strategy of mediating the difficulties was applied in (Wolf et al., 2011) where the edges that correspond to the instrument are isolated from the miscellaneous background gradients by filtering for gradients that were consistent with the known insertion point or *trocar* of the instrument. As the position of the insertion point with respect to the camera can be computed a priori, the possible pose of the camera reduces to a 4 DOF transformation where the (x, y) translational DOFs are eliminated. This allows the range of possible gradient orientations for the shaft edges to be constrained to a reasonably small range. Although this method yielded reasonable results, the segmentation steps involve thresholding and noise removal to yield workable images which suggest that the method may not be hugely robust to particularly noisy images.

3.2.3. Ad-hoc Detection

Finally, a third category of methods exists which does not rely on the creation of data-driven models or optimization designs.

Methodological Overview. Ad-hoc methods most generally rely on the use of low-level image processing techniques such as thresholding or local extremum identification to detect surgical tools and estimate their pose parameters. Transformations of the input image into black and white masks, easily derived from traditional feature channels, represent the favored type of data input due to ensuing processing easiness. Usually, approaches are built as multiple step frameworks performed in an iterative fashion, each step improving pose estimation accuracy. After enhancing input data by applying de-noising techniques, a tool overall location within the image can be identified, then enabling to refine pose parameters by estimating specific tool landmark location (e.g. tip, center) or tool orientation.

Literature Approaches. Leaving aside optional techniques applied for image de-noising or enhancement, the overall tool location is the first pose parameter needing to be estimated. This estimation usually revolves around the identification of the biggest blob regions in the image as no tool models are being learned. Region growing algorithms (Speidel et al., 2008, 2014), simple clustering of points with minimal values in standard image feature channels (e.g. saturation and range (Haase et al., 2013)) have been proposed. On top of semantic labelling maps, a weighted averaging technique has been employed to identify the location of different parts of a single tool (Sznitman et al., 2014). Additionally, combinations of blobs and edge detections have also been proposed (Alsheakhali et al., 2015) whereby blobs are used to filter the output of a Hough line detector. However, blob region estimation techniques heavily suffer in terms of robustness and reproducibility from the use of hard-coded criteria such as the minimal size in pixels for a region, or thresholds to eliminate false candidate regions.
From a correct tool location estimation, the focus for upcoming steps is brought toward the estimation of remaining pose parameters. Assuming the tool-tip to be located somewhere along the center-line in-between tool shaft lines, the tip position can be identified as the point with the highest gradient which represents the transition from background to tool (Haase et al., 2013; Zhou & Payandeh, 2014). Similarly, a color criterion can be used to identify the transition between shaft and tip along the tool main

axis (Speidel et al., 2008). In (Speidel et al., 2014), the maximal-distance points within each obtained blob region are computed and the one farther from image boundaries is considered being the tool-tip.

Regarding orientation or shaft width estimation, ad-hoc methods hypothesize surgical tools to be rigid and tubular. Line patterns, usually representing center or shaft-border lines of the instrument, are hence specifically searched for. Edges of interest have been computed using Sobel filtering techniques (Haase et al., 2013) or by thinning previously obtained blob regions up to their main axis using a skeletonization approach (Speidel et al., 2008). Then, existing lines are often retrieved by applying the Hough transform over resulting edges. Alternatively, a RANSAC approach has also been used by fitting a line over the elements of a region of interest (Sznitman et al., 2014).

Ad-hoc methods can be assimilated as low-level and naive techniques that are not to be favored in general as relying on too many assumptions. In addition, empirically defined thresholds are not easily transferable to other surgical contexts or to detect other surgical tools. Moreover, such approaches can be easily disrupted by a number of factors such as image noise, tool occlusion, and lighting variations.

3.3. Prior Knowledge

In order to constrain the detection search space, thus facilitating the task, many approaches rely on a set of assumptions, or prior knowledge (fourth set of columns table 2). Such knowledge having different forms and aspects, we chose four categories for its representation: assumption over tool shape and location within the image, user assistance, and robot kinematics. In the following, an overview of each category and some examples are provided.

3.3.1. Tool Shape Constraints

Assumptions over the shape have been widely employed to design detectors. Low-level considerations regarding tools as simple tubular shapes (Speidel et al., 2008; Sznitman et al., 2014) or solid cylinders with a tip alongside the center-line (Allan et al., 2013; Burschka et al., 2005; Haase et al., 2013; Zhou & Payandeh, 2014) have been used. Similarly, rough estimates of a tool shape have been expressed, either being represented by two edges symmetrically spaced from an axis-line containing the tip (Voros et al., 2007) or by two parallel side segments and a tip lying in-between (McKenna et al., 2005). On the other hand, highly detailed shapes with joint configurations can be leveraged from a rendering software (Pezzementi et al., 2009; Reiter et al., 2012a).

3.3.2. Tool Location Constraints

The second most common type of assumption relates the known intersection between the surgical tools and image boundaries. Surgical instruments are expected to enter the scene from image boundaries (Allan et al., 2013; Haase et al., 2013; McKenna et al., 2005), thus being visible on image edges (Sznitman et al., 2013). Sometimes the constraint is expressed within the processing algorithm, where the corresponding initialization is performed by looking exclusively at image border areas (Speidel et al., 2006), or by choosing candidate regions close to image boundaries (Doignon et al., 2004).

3.3.3. User Assistance

Instead of relying on generic assumptions over a tool shape or its location within the image, some methods request manual help from the user. For minimally invasive surgeries, the knowledge of the instrument insertion point in the cavity greatly constraints the search space to a limited beam. The insertion point can be selected by the surgeon using a vocal mouse (Voros et al., 2007) or after computation requiring manual selection of 2D instrument boundaries in a sequence of images (Wolf et al., 2011). In case of online learning algorithms or tracking by initialisation, a user may also have to indicate to the system which image portions are containing surgical tools needing to be subsequently identified (Reiter & Allen, 2010; Sznitman et al., 2012).

3.3.4. Robot kinematics

The cable driven kinematics of daVinci robots leads to relatively inaccurate tool pose estimations supplied by internal encoders. At the same time, such inaccurate poses can be seen at good estimates to constraint the search. Robot kinematics data have been used as input to the detector (Burschka et al., 2005), or to render on-the-fly tool models with a limited set of joint configurations (i.e. different tool poses) (Reiter et al., 2012a). Lastly, robot kinematics can be used in a post-processing manner to reject erroneous detections or fill the gap of missed ones (Reiter et al., 2012c).

3.4. Temporal Tracking

An important component of a detection system is to link the measurements temporally to obtain a smoother trajectory and also handle situations when the instrument may be heavily occluded. Normally the instrument pose parameters are represented with a *state vector* which is transformed from frame to frame to obtain a consistent measurement.

3.4.1. Sequential Bayesian Filters

The most simple filter of this type is the Kalman Filter. This provides an optimal estimate of the state vector at time t given a measurement of it at t and the prior state at time $t-1$. It makes the assumption that the distribution over the world state is normal and that the model which maps measurements to state vectors is linear with normal additive noise (Prince, 2012).

Burschka et al. (Burschka et al., 2005) made use of a Kalman filter to combine measurements from a da Vinci robotic control system with visual measurements as part of a robotic servoing system. To avoid failed visual observations from corrupting their state estimate, they threshold their visual observation confidence and only make use of the measurement prediction from the Kalman filter when the threshold inaccuracy is exceeded. (Zhou & Payandeh, 2014) also made use of a Kalman filter as well as the extended Kalman filter which allows the use of non-linear models such as polar coordinates for the pose parameters.

3.4.2. Particle Filters

Particle filters represent the probability function over the state with a set of particles which are evolved through time by a particular model of the system. A well known particle filtering method is the *Condensation* algorithm (Isard & Blake, 1998). Each particle represents one estimate of the system state and at each timestep it is projected through a, possibly non-linear, state transition function giving a new estimate of the system state. This estimate is then evaluated giving a probability of its accuracy. A new set of particles can then be estimated by resampling from this new distribution.

The Condensation algorithm is popular in surgical instrument tracking due to its ability to track through occlusion which is achieved by maintaining multimodal state distributions. (Wolf et al., 2011) make use of the this algorithm to estimate the pose of medical instruments as part of a framework that incorporated geometrical constraints from the known insertion point. They divide a semi-sphere around the surgical cavity into panels, where each panel represents a 3D vector projecting through it from the known insertion point. They score the state estimate of each particle by backprojecting the instrument into the image and comparing the pose estimate. The instrument tip is found using an Otsu's threshold along the projected central axis of the estimated instrument. Speidel et al.(Speidel et al., 2006) also made use of the Condensation tracker for localizing instruments in medical images.

An alternative particle approach of (Hue et al., 2000) was introduced to the medical field in (Speidel et al., 2014). Although single particle filters can approximate multi-modal distributions, they tend to approach uni-modal distributions over time. To counter this, the proposed method tracks different regions of the target by a multi-object particle filter which represents the object state as a concatenation of several configurations simultaneously.

3.4.3. Initialisation

Several additional methods (Kumar et al., 2013b; Reiter & Allen, 2010; Richa et al., 2011a; Sznitman et al., 2012) of instrument detection take a much simpler method of fusing temporal information by simply initialising their search for the next frame's detection at the location of the previous frame's detection. When this initialisation must be performed by the user, a ✚is indicated in Table 2).

3.5. General Optimization Strategies

For integration in real-time in-vivo medical applications, highly accurate tool detectors are key, but not at the detriment of the processing speed. Finding the optimal speed versus accuracy trade-off is usually difficult. Computer hardware specifications, code optimization, use of parallel computing, and image resolution have a significant impact on speed performances. As a consequence of this variability, we chose not to integrate this information into the table as a direct speed comparison between detectors would not yield much sense. However, we propose to report interesting optimization strategies mentioned by authors to increase the computational speed.

3.5.1. Search Space Reduction

A naive way for performing detection speed-up is to reduce the search space range or specific pose parameter ranges. For detectors based upon a sliding window approach, the most popular ad-hoc optimization implementation is to reduce the number of pixels to process. It can be achieved by performing spatial down-sampling (e.g. factor 2 to 4) over input images (Pezzementi et al., 2009; Voros et al., 2007), or by processing every fourth line and column (Speidel et al., 2006). Similarly, when dealing with video inputs not every frame needs to be processed, assuming a recording speed between 25 and 30 Hz, because of the limited motion of surgical tools within consecutive frames. Speidel et al. (Speidel et al., 2013) proposed to process every fifth frame, while Reiter et al. (Reiter & Allen, 2010) processed every third frame.

Finally, for brute-force approaches requiring to process large amounts of pose-specific models, a coarse-to-fine approach can remedy the huge processing time issue (Reiter et al., 2012a). For example Sznitman et al. (Sznitman et al., 2012) used a Gaussian Pyramid approach where the detector exhaustively visits every location in the Pyramid. Alternatively, using an inaccurate external tracking system such as robot kinematics can be used to constrain the brute force search (Reiter et al., 2012b).

3.5.2. Algorithmic Reduction

The feature extraction process, shared by most dicriminative and generative approaches, requires an important processing time and its optimization can greatly increase speed. An efficient way of computing channel features, called *integral channel features* (Dollár et al., 2009), has been proposed where sums over local rectangular regions are extracted from each image channel. Integral channel features combine the richness and diversity of information from use of image channels with the computational efficiency of the Viola and Jones detection framework. A number of papers have utilized integral channel features for different applications such as object recognition (Laptev, 2006; Tu, 2005), pedestrian detection (Dollár et al., 2007), edge detection (Dollar et al., 2006), local region matching (Babenko et al., 2007), brain anatomical structure

segmentation (Tu et al., 2008), and surgical tool detection (Bouget et al., 2015).
When the detection problem can be expressed as a differentiable function, as is typical in generative methods, gradient based methods such as steepest descent (Allan et al., 2013) and variants of second order methods such as Gauss Newton (Richa et al., 2011b) have been used to search in a locally optimal manner. When derivatives are not available, global search can be avoided with the Nelder Mead simplex method (Pezzementi et al., 2009). Finally, for discriminative approaches using decision forests, limiting cascade classifiers parameters, especially the tree length and depth, to a minimum has been proposed to increase computational speed. Early stopping scheme (Sznitman et al., 2014) or manual limitation (Allan et al., 2013) have also been proposed towards this effect.

4. Validation Methodology

In order to quantify surgical tool detector performance and perform rankings in a realistic, unbiased, and informative manner, a proper and well-defined validation methodology is required. To do so, we propose to investigate existing tool detection validation methodologies through their specification phase (high-level) and computation phase (low-level). In the former, we explore the objective, the validation type and the model validation. In the latter, we examine validation criterion and its estimation by focusing on figures of merit, validation metrics, and normalization steps. Most of previous studies performing *validation*, as specified in (Jannin et al., 2006), the term of *validation* is used to refer to the methodology and relative components. In the following, both categories are presented in details and Table 3 provides an overview of collected information.

4.1. Specification Phase

The specification phase of the assessment methodology defines the conditions in which the assessment is being performed with a clearly formulated assessment objective and type.

4.1.1. Objective

In each and every previous study, the *validation* performance has been assessed, normally through detection quality (Kumar et al., 2013b; Sznitman et al., 2012; Wolf et al., 2011; Voros et al., 2007) which is more frequently analysed than the quality of the tracking component (Reiter & Allen, 2010; Sznitman et al., 2013). Verification and evaluation are less frequently assessed than validation. Verification has been used to obtain insights into method strengths and weaknesses (Sznitman et al., 2013). Evaluation has been performed for practical value demonstration in an eye surgery proximity detection task context (Richa et al., 2011a).
The vast majority of assessment have been carried out in the 2-dimensional space, and only a few in the 3-dimensional one (Burschka et al., 2005; Haase et al., 2013; Wolf et al., 2011).

4.1.2. Validation Type

Commonly, a study can be assessed in two different ways: qualitatively and quantitatively. The former returns insights after visual observation of a phenomena. The latter corresponds to a systematic empirical investigation of observable phenomena through the computation of statistical or numerical values.
Most studies report detector performance in a quantitative way, explained in details in the following section. Regarding qualitative assessment, it can be expressed through numerous variants such as images with overlaid detection results (Speidel et al., 2006) or plots showing the evolution of one parameter within the image referential (Richa et al., 2011a).

4.1.3. Model Validation

The model validation strategy is crucial for assessing the external validity of the model, which demonstrates the extent to which the results of a study can be generalized to other surgical contexts or tools. In a prediction problem, the model training is performed over a data-set of known data, and the model is tested against a data-set of unknown data.
Standard data-set splitting has been used, where either the first half of every sequence is collected into the train split and the other halves represent the test split (Sznitman et al., 2012, 2013, 2014) or with randomly balanced train and test sets (Bouget et al., 2015). More robust validations, using a cross-validation strategy, have been employed; in a leave-one-out manner (Sznitman et al., 2012), or in a 10-fold way (Kumar et al., 2013b). In some cases, the data-set separation into train/test sets is unclear and the same images may appear in both sets (Speidel et al., 2006, 2008). In cases of online learning (Reiter & Allen, 2010) or in methodologies where no learning phase is implemented (Voros et al., 2007) there is no need to separate train and test data.

4.2. Computation Phase

The computation phase of the validation methodology expresses how the estimation of a validation criterion is being performed. Three elements describe the quantification of a validation criterion: a comparison method (i.e. metric), information on which the computation is performed (i.e. reference), and a figure of merit (i.e. quality index) (Jannin et al., 2006).

4.2.1. Criterion

A validation criterion aims at characterizing different properties of a method such as its accuracy, precision, robustness, or reliability. This information is not reported in the table as every study has exclusively focussed on both accuracy and precision. Some attempts have been made

Table 3:
Validation methodology: overview of methodologies used in the literature to validate surgical tool detection methods. **Type**: Verification (Verif.), Validation (Valid.), Evaluation (Eval.). **Error statistics**: Mean (M), Standard Deviation (S), Order statistics (O). **Standard Performance Measures (SPM)**: Recall (R), Precision (P), Probability of error (PE), Accuracy (A), True Positive Rate (TPR), False Positive Rate (FPR), False Positives Per Image (FPPI). **Metric**: Intersection Over Union (IOU), Number of Consecutive Frames (NCF).

	Specification							Computation										
	Objective			Model		Type		FOM			Normalization				Metric			
	Type	2D	3D	Split	Cross-Validation	Qualitative	Quantitative	Error stats.	SPM	Duration	Landmark	Orientation	Bounding Box	Pixels	Distance	IOU	NCF	Visual
(Allan et al., 2013)	Verif.		✔				✔	M,S			✔	✔			✔			
	Valid.		✔			✔					✔	✔						✔
	Valid.	✔			✔		✔		R,P,PE					✔		✔		
	Valid.		✔		✔		✔		R,P,PE				✔			✔		
	Valid.		✔				✔				End				✔			
(Allan et al., 2014)	Valid.		✔				✔	M,S				✔			✔			
(Allan et al., 2015)	Valid.		✔				✔	M,S				✔			✔			
(Alsheakhali et al., 2015)	Valid.	✔					✔		TPR		Tip				✔			
(Bouget et al., 2015)	Valid.	✔		✔			✔		FPPI					✔		✔		
		✔		✔			✔		R		Tip	✔			✔			
(Burschka et al., 2005)	Valid.	✔				✔					✔							✔
			✔			✔					Tip							✔
(Cano et al., 2008)	Valid.		✔				✔	M,S	A		Tip				✔			
(Doignon et al., 2005)	Valid.	✔				✔							✔					✔
(Doignon et al., 2007)	Valid.		✔			✔					✔	✔						✔
(Haase et al., 2013)	Valid.	✔					✔	M,O			Tip				✔			
			✔				✔	M,O			Tip				✔			
(Kumar et al., 2013b)	Valid.	✔			✔		✔		A				✔			✔		
(Li et al., 2014)	Valid.	✔					✔		R		Tip				✔			
		✔		✔			✔	M									✔	
(McKenna et al., 2005)	Valid.	✔					✔	M				✔			✔			
(Pezzementi et al., 2009)	Valid.	✔					✔		P,R,PE					✔		✔		
		✔					✔	M			✔	✔			✔			
(Reiter & Allen, 2010)	Valid.	✔					✔			✔							✔	
		✔				✔					End							✔
(Reiter et al., 2012c)	Valid.	✔		✔			✔		A		Center							✔
(Reiter et al., 2012a)	Valid.	✔					✔		R		✔	✔						✔
(Richa et al., 2011a)	Valid.	✔				✔					✔							✔
	Eval.	✔					✔				Tip				✔			
(Rieke et al., 2015)	Valid.	✔		✔			✔		R		Tip				✔			
					✔		✔		PCP		✔				✔			
(Speidel et al., 2006)	Valid.	✔				✔					✔							✔
(Speidel et al., 2008)	Valid.	✔					✔	M			Tip				✔			
		✔					✔				Tip				✔			
(Speidel et al., 2014)	Valid.	✔					✔		R,P					✔		✔		
		✔					✔	M			Tip				✔			
(Sznitman et al., 2012)	Valid.	✔		✔			✔		R		Tip				✔			
		✔			✔		✔		R		Tip				✔			
		✔		✔			✔	M									✔	
(Sznitman et al., 2013)	Verif.	✔				✔												✔
	Valid.	✔		✔			✔		TPR,FPR,P		Tip				✔			
	Valid.	✔		✔			✔	M,S			Tip	✔			✔			
	Valid.	✔		✔			✔				Tip						✔	
(Sznitman et al., 2014)	Valid.	✔		✔			✔		R		Center				✔			
		✔		✔			✔	M,S			✔	✔			✔			
		✔		✔			✔	M,S			Center				✔			
(Voros et al., 2007)	Valid.	✔				✔					Tip							✔
		✔					✔	M			Tip				✔			
(Wolf et al., 2011)	Valid.	✔					✔	M,S				✔			✔			
			✔				✔	M,S				✔			✔			
		✔					✔	M			Tip				✔			
(Zhou & Payandeh, 2014)	Valid.	✔				✔		M	TPR		Tip	✔						

to retrospectively study the robustness, but it was not the intended objective for the study and as such can not be considered as the validation criterion. Using in-vivo data, with full clinical realism but no control, it is difficult to target the validation of either robustness or reliability.

4.2.2. Normalization

The first element necessary to perform the validation is the reference, also called Gold Standard or Ground Truth. This element is supposed to represent the absolute truth and is meant to be used as basis for comparison with the results of a method. In general, the validation is rarely directly performed with the whole reference, but rather with some sub-elements composing the reference. For surgical tool detection validation, information contained in the reference are for example the tool location, its orientation or its tip position. As such, a normalization step is usually performed prior to the validation in order to transform the reference and detection results into an equivalent representation.

The favored reference, used in every study, is a landmark on the tool: either the tip (Speidel et al., 2008; Sznitman et al., 2012, 2013; Voros et al., 2007), the center (Reiter et al., 2012c; Sznitman et al., 2014), or the end (Reiter & Allen, 2010). However in several studies (Pezzementi et al., 2009; Speidel et al., 2006) the overall tool pose is not limited to a specific landmark.

The second most common reference is the orientation of the tool shaft (McKenna et al., 2005; Wolf et al., 2011). While few works exploited tool bounding boxes, either as direct reference in Kumar et al. (Kumar et al., 2013b), or by deriving pixel-wise tool label maps (Pezzementi et al., 2009; Speidel et al., 2014).

4.2.3. Metrics

The metric is a comparison function measuring a distance between the normalized results of the method and the corresponding normalized reference. Previously used metrics can be regrouped in four categories: metric distance, Intersection Over Union (IOU), Number of Consecutive Frames (NCF) and visual criterion.

The metric distance, often computed as the Euclidean distance, is favored when the reference is a single value (e.g. tool orientation) or a point (e.g. tip position). The metric, usually used for simple computation, can also be used in a thresholding fashion to separate true positive from false positive detections. For example, a detection is considered accurate for a distance error under 10 pixels (Sznitman et al., 2013), or recall values are reported following an evolving distance threshold (Sznitman et al., 2012, 2014).

The Intersection Over Union criterion metric has been used with the standard 50% overlap threshold between bounding boxes in (Kumar et al., 2013b) and with a 25% overlap value in (Bouget et al., 2015). A variant has also been proposed, where the criterion is not employed with bounding boxes but rather over the full image in a pixelwise fashion (Pezzementi et al., 2009; Speidel et al., 2014).

Both aforementioned metrics operate towards spatial detection performance. A metric dedicated to the tracking aspect has been proposed by (Reiter & Allen, 2010) and (Sznitman et al., 2012). The Number of Consecutive Frames (NCF) until the tracker loses the tracked detection is considered. Finally, observer-based visual metrics rely on the observer to evaluate the quality, for example where the tool center-line must be within the tool shaft (Reiter et al., 2012c), or where the tool-tip location must be accurate with according joint configurations (Reiter et al., 2012a).

4.2.4. Figures of Merit

The figure of merit, or quality index, is used to obtain a statistical measure of the distribution of local discrepancies computed using the validation metric (Jannin et al., 2006). Three figure of merit types have been identified: standard statistics, Standard Performance Measures (SPM) (Makhoul et al., 1999), and duration.

Standard statistics relate to error computation, most of the time, of pixel values. Examples are mean (M) error (McKenna et al., 2005; Wolf et al., 2011), standard deviation (S) of the error (Kumar et al., 2013b; Sznitman et al., 2014; Speidel et al., 2013), or order statistics (O) of the error (Haase et al., 2013).

Standard performance measures, also expressed as information retrieval metrics, cover the calculation of true positive, true negative, false positive, false negative, and all entailing measurements such as recall (R) (Sznitman et al., 2012, 2014), precision (P) (Allan et al., 2013), accuracy (A) (Kumar et al., 2013b) and probability of error (PE) (Pezzementi et al., 2009). In cases of by-parts validation (Rieke et al., 2015), the strict Percentage of Correct Parts (PCP (Ferrari et al., 2008)) addressing the length of the connected joints of the tool can be used.

The duration has only been used once to report an elapsed time in seconds (Reiter & Allen, 2010).

5. Alternative Detection Methods

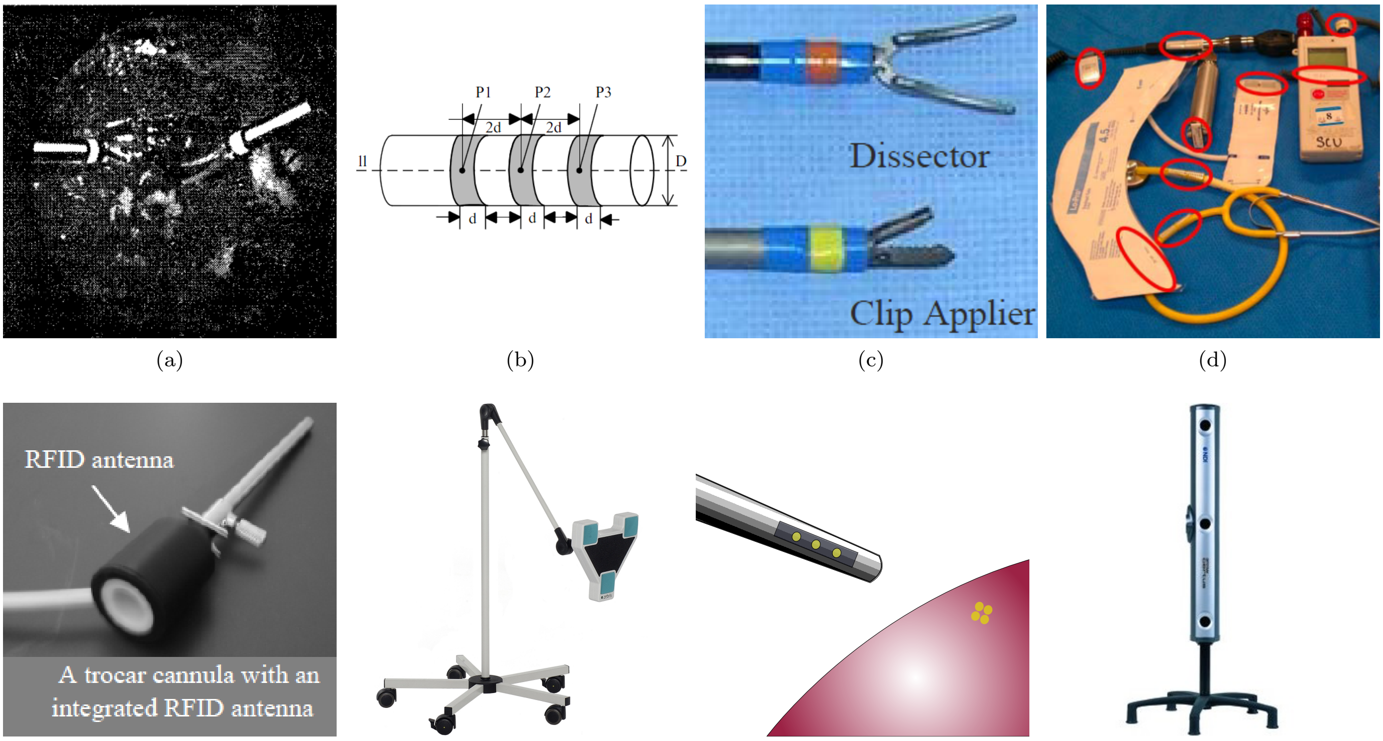
The instrument detection and tracking methods suggested so far in the review cover methodologies that make no modification to the design of the instruments or the surgical workflow. This is generally seen as a desirable quality (Stoyanov, 2012) as the clinical translation of this type of method is comparatively straightforward as modifications have sterilization, legal and installation challenges. However, as illustrated throughout this review there are many significant challenges around markerless image processing methods from visual occlusion, lighting and inaccurate depth estimation which motivates the use of of alternative detection methods accepting the modification of the design as a necessary complication.

- Color and Shape Markers: Early approaches involved coating the instrument in a color easily separable from the tissue background (Ko et al., 2005; Wei et al., 1997; Tonet et al., 2007) (see Fig. 7c). To select the most suitable color, a distribution analysis based on RGB or HSV colorspaces is often performed. As for marker identification, simple thresholding techniques in corresponding colorspace image channels only are necessary. More recently, biocompatible color makers have been evoked (Bouarfa et al., 2012). Sometimes a distinguishable shape is also enforced, such as multiple-parts black markers (Casals et al., 1996; Zhang & Payandeh, 2002) (see Figs. 7a and 7b).

- Optical Trackers: One non-vision approach to tracking the pose of medical instruments is to use optical markers, reflecting or emitting a non-visible wavelength of light (e.g. infra-red). By measuring the location of several of these markers with a stereo camera system, rigid transformations between the marker coordinate systems and the camera coordinate system can be estimated. Common devicess include the Optotrak ® system of NDI [7] (see Fig. 7h) used to track instruments as part of a larger augmented reality framework MEDIASSIST (Speidel et al., 2008). Similarly, Krupa et al. (Krupa et al., 2003) attached LED markers and an infra-red laser to a laparoscopic instrument shaft and isolated the laser projection through high-pass filtering (see Fig. 7g).

- Acoustic Trackers: Typical acoustical tracking systems make use of time-of-flight measurements of ultrasonic sound waves to deduce the position and orientation of the target object. Zebris Medical GmbH

[7] http://www.ndigital.com/



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 7: External markers and sensors used as part of alternative surgical tool detection methods. (a) Shape marker (Casals et al., 1996), (b) Shape marker (Zhang & Payandeh, 2002), (c) Color marker (Ko et al., 2005), (d) RFID tags (Parlak et al., 2011), (e) RFID system (Miyawaki et al., 2009), (f) Acoustic tracker (Chmarra et al., 2007), (g) Optical device (Krupa et al., 2003), (h) Optical tracker (Speidel et al., 2008).

developed the CMS10 system (see Fig. 7f) which tracks the 3D coordinates of miniature ultrasonic transmitters relative to three microphones (Chmarra et al., 2007).

- Mechanical Trackers: Modern robotic surgical systems, such as the da Vinci ® of Intuitive Surgical [8], are typically equipped with a series of cable driven arms. By measuring the current state of each cable in its arms, the system is able to make an estimate of the location and orientation of each component of the arm. Although providing accuracy estimates considering the pertaining challenges, results are not accurate enough to support the type of guidance and navigation required.

- Electromagnetic Trackers: This tracking method is based on the currents generated in small wire coils positioned in a magnetic field. One popular system produced by Immersion [9] for laparoscopic simulations is known as the Laparoscopic Surgical Workstation ®. This records the position of laparoscopic instruments with electromagnetic transducers mounted in a gimbal (Chmarra et al., 2007).

- Shape Sensors: Another non-vision approach to instrument localization is represented by shape sensors, such as DSS 4300® by Luna Innovations [10]. Shape sensors are built as kinematic chains of shape sensing segments where each segment detects local strain using optical properties such as fibre Bragg gratings or Raleigh scatter patterns (Duncan & Froggatt, 2010; Prisco, 2010).

- RFID Systems: Non-vision approach consisting in emitting and receiving radio-frequencies in combination with surgical instruments preemptively equipped with RFID tags. Often, the technology has been employed for surgical event recognition more than for accurate 2D or 3D pose estimation (Parlak et al., 2011; Bardram et al., 2011) (see Fig. 7d). Tool detection has either been performed using wireless palm-based RFID readers, or through a RFID antenna mounted on the operating room ceiling. A RFID system specific for tool detection during minimally invasive surgeries has also been proposed where RFID readers are mounted directly on trocar cannulas (Miyawaki et al., 2009) (see Fig. 7e). However, issues might arise especially because of interference, reflection, and shielding (Van Der Togt et al., 2008; Houliston et al., 2009; Christe et al., 2008).

[8] http://www.intuitivesurgical.com/
[9] http://www.immersion.com/
[10] http://lunainc.com/

6. Discussion

Image-based surgical tool detection and tracking methods have been studied for almost two decades and have made marked progress in conjunction to advances in general object detection within the computer vision community. We expect the field to grow and have increased importance because surgery as a field is fully committed to the MIS paradigm which inherently relies on cameras and imaging devices. In this paper, we have reviewed the main lines of exploration so far in image-based detection and tracking and we now attempt to identify the key challenges and areas for development.

6.1. Data for Surgical Tool Detection and Tracking

The lack of available high quality data to use for development as well as for algorithm validation, comparison and benchmarking is a significant bottleneck for the tool detection and tracking community. The value of benchmarks is well recognised (Benenson et al., 2014) because without the practice individual papers simply show a narrow view over the state-of-the-art and inhibit both authors and reviewers to easily access the merit of new publications. The effect of data availability and benchmarking has been demonstrated to act as a catalyst to high quality development and progress in various fields in computer vision (e.g. computational stereo with Middlebury Stereo Vision dataset or in pedestrian detection with the Caltech-USA dataset). However, recently authors have begun to make data available online (e.g. (Bouget et al., 2015; Kumar et al., 2013b; Sznitman et al., 2014)). A notable development in the community has been the sub-challenge on instrument segmentation and tracking, as part of the endoscopic vision challenge at MICCAI 2015 [11], where annotated datasets for both detection and tracking have been provided for laparoscopic and robotic MIS instruments.

A possible explanation to the lack of established validation and testing data is that acquisition of surgical video and images was traditionally a difficult, labour intensive and procedure dependent task. Different surgical systems required a variety of video acquisition sources for different video standards and custom and often non-constant calibration. Regulatory issues need to be considered to record the data and furthermore to allow access to it. However, with technology standardisation and growing maturity in the CAI field data is becoming more tangible and datasets have started becoming available for MIS, ophthalmology and neuro-surgery. Additionally, stimulated by surgical training and digitisation of educational tools surgical videos have started becoming available for surgical trainees and to explain new surgical techniques. It is important for this practice to continue in order to generate sufficiently large datasets (some surgeries are rare) for a range of procedures across different hospitals and equipment.

When gathering data, automatic or random selection processes must be favored over manual ones in order to reduce selection bias. As such, video segment selection for subsequent split into images is preferable to stand-alone images selection (Dollár et al., 2011). Aside from reducing bias, selecting video sequences also provides an additional pool of information to process. Temporal correspondences between tool instances within a sequence enable trajectory analysis, and temporal features computation (e.g. optical flow). Not to mention the inability for tracking systems to exploit a data-set made of stand-alone images only. Similarly, depending on camera hardware used for the recording, either monocular or stereoscopic images are obtained. While the former category is represented in the vast majority of data-sets, only the latter provides an access to additional features such as depth maps. As a result, depending on the type of data constituting the data-set, some features may not be accessible for computation.

Instrument tracking video data can also be acquired in simulated or phantom environments. In such cases, clinical realism and visual fidelity is usually sacrificed for the ability to validate or test comprehensively subject to specific conditions such as illumination changes and occlusions. However, only in-vivo videos capture the full complexity of clinical data realism depending on surgical fields and tool appearance variations. Ideally datasets should be comprehensive enough to allow for conditions such as occlusions, illumination variations, or motion blur with sufficient data sizes. Strategies to access sub-sets covering specific ranges of challenging conditions by relying on the information gathered through the annotation process and meta-data are also highly desirable.

Data annotations are crucial for reference during validation of algorithms and benchmarking but can also be used for machine learning approaches to instrument detection and tracking. As user interaction during annotation is a variable process, data control for annotation is a crucial issue that has complex trade-offs with cost and time implications. Recently, crowd-sourcing solutions have been proposed (Maier-Hein et al., 2014) and shown to provide robust data annotation with interesting statistics on the limited variability in labelling between expert surgeons and large groups of novices. In addition to this, commercial solutions have emerged to simplify the interaction between researchers and the annotating volunteers[12]. The complexity of annotations can vary between polygon bounding boxes or hulls (Bouget et al., 2015) and/or tool tip locations or pixel wise segmentations (Speidel et al., 2008). Particular to surgical instruments is the difficulty in annotating in-plane rotations or the pose of articulated instruments. Additionally, for certain methodologies it may be useful to have very detailed mark ups and meta data on occlusions, blur, blood, or complex lighting effects and this is an area that requires further efforts.

[11] http://endovis.grand-challenge.org/

[12] http://pallas-ludens.com/

An exciting possibility for data generation is though robotic surgical systems such as the daVinci and RAVEN systems or optical trackers which provide a solution for automatic collection of annotated 3D data (Allan et al., 2014, 2015). However, inaccuracies in measurements arise from camera and hand-eye calibration, the multiple joint encoders, and the lack of translation to non robotic instruments. Due to the challenges in obtaining the ground truth with a highly accurate method, annotating the data often has to be corrected manually. If this manual procedure can be overcome, the robotic approach can potentially provide powerful and rich datasets.

6.2. Methods for Tool Detection and Tracking

Regarding image-based object detection techniques, performance is determined by two key factors: the chosen feature representation and the learning algorithm used during the pose estimation process (Dollár et al., 2009). Steady improvements have been made in both learning and feature design by taking advantage of a faster progress of those areas in many other computer vision departments.

6.2.1. Feature Representation

Variants or combinations of more than eight different feature families can be extracted from images, hence causing the feature representation to be a highly versatile component. While spatial features (e.g. SIFT, HOG) have been implemented extensively, many newer features such as stereo and temporal cues (e.g. depth map, optical flow) remain yet to be fully exploited. Despite its significant popularity in many object detection areas (e.g. face, pedestrian), the HOG feature representation is far less popular for tool detection where color features have been heavily favored. Choosing the appropriate color space can be driven by the object to detect, and HSV has been stated to be better suited than RGB for surgical tool detection (Speidel et al., 2006). A separation between chromaticity and luminance components is offered by this color space, therefore bringing more robustness to illumination changes. Instead of relying on standard feature representations, many tool detectors have been favoring semantic labels as features. While many different classes can be modelled towards object detection in daily life images (e.g. road, grass, building, or sky), their number is highly reduced for surgical tool detection. Indeed, given the nature of surgical instruments being gray metallic objects, usually only two classes are necessary: one to model tool pixels and one to model background pixels. Evidently, the more classes the less accurate semantic labels maps, as differences between classes will be more and more subtle. Leveraging such contextual features seems to be a promising solution for surgical tool detection and more similar information could be used in the future (Galleguillos & Belongie, 2010).

An adequate feature representation is generally hard to find and the most popular approach for quality improvement is to increase and diversify the features computed over the input image (Benenson et al., 2014). Having richer and higher dimensional representations tends to ease the classification task, enabling improved results. However, developing a more profound understanding of what makes good features good and how to design them is still needed. Up to now, improvements were made through extensive trial and error. Additionally, the integral feature channel representation is being more and more used, also for surgical tool detection (Bouget et al., 2015; Reiter et al., 2012c). This representation offers easier and faster access to feature values over different configurations of rectangles, compared to standard feature channels.

6.2.2. Pose Estimation

The main objective for every learning-based approach is to generalize well from train to test set, which is usually described as the model accuracy versus generalization trade-off (Fawcett, 2006). Over-fitting to the training set involves learning model parameters which explain noise or insignificant details which can be understood as learning something by heart. The opposite situation occurs when the model has insufficiently complexity to learn the relationship between inputs and outputs causing low test set accuracy. Models that can generalize well are compulsory for detectors to be able to identify surgical tools throughout various surgical procedures coming with slight to moderate background and tool appearance variations. Interestingly enough, data-driven or model-based approaches (i.e. generative and discriminative) have not been the only methods used for surgical tool detection where many ad-hoc threshold-based approaches have been considered. This type of strategy is no longer used in more advanced object detection fields as it induces too much bias and not enough reproducibility in the way thresholds are defined. According to Sznitman et al. (Sznitman et al., 2014), building classifiers to evaluate the presence of surgical instruments appear to be the most promising solution for both in-vivo detection and tracking, and as such data-driven learning strategies should always be favored.

For discriminative detection, many learning paradigms exist which can be assimilated to one of the three representative families: decision forests, deformable part models, and deep networks. As of now, surgical tool detection studies have yet to make use of them all, generally relying on the oldest paradigms such as SVM and decision forests classifiers. Moreover, each family has been shown to provide extremely results close to the others indicating the choice of the learning paradigm not to be a dominant one (Benenson et al., 2014). For all the aforementioned model learning strategy, accurate training data is mandatory in order to build a proper model. An alternative, for tools belonging to the category of robotic devices, is to use a robot renderer with a CAD model in order to generate tool templates according to specific kinematic joint configurations (Reiter et al., 2012a). This is stated to be desirable because collecting training data becomes easier than if it had to come from videos, thus enabling larger collection with less effort. Advantages of this type of data generation have been shown successfully in (Shotton et al., 2013). However, choosing appropriate object parts to model can also prove challenging, particularly due to occlusion from other instruments, tissue and from the field of view. For surgical tools, modelling the tip region is the most viable tactic as it is the most characteristic landmark for tool differentiation and is the most likely component to be in view, relative to the tool end or tool body. However, tool tips can be cumbersome to model when made of many parts, which is the case for articulated surgical instruments. In general, creating full surgical tools models is often unnecessary due to large sections of the instrument being out-of-view. However, if starting from fully modelled surgical tools, previous works on occlusion handling could prove useful to better derive sub-models (Mathias et al., 2013). The future of discriminative approaches probably lies in the use of Convolutional Neural Networks, which are already on a growing trend in other fields of object detection. The main hurdle is the required amount of training data, currently too high with respect to the validation data-sets size described in this paper. Nonetheless, the use of already pre-trained networks has shown promising results and should be further investigated in a medical context.

Generative methods are at least as popular if not more popular than discriminative methods in instrument detection as they provide several notable advantages in a surgical context. The primary advantage lies in the modelling processing, which for a generative model is often closely coupled with the physical process through which the image features are formed. This allows the model to be adapted to incorporate prior knowledge or reason about information from different sources, such as additional tracking markers. In addition to this, generative methods have a much lower requirement on training data, sometimes requiring only a few images to estimate the required parameters compared with discriminative methods which often require huge training databases. Future advances in generative methods can be achieved by combining different edge, region and point based methods within single frameworks which enable the different strengths of each type of feature to balance failure cases in an individual feature. Achieving this however raises important challenges around how fast and reliable optimization can be performed. Three main feature representations in generative methods have been covered: points, regions and edges with regions as the most popular by far with more than half of the referenced papers using this type of method. This fact is unsurprising given the relative simplicity of using region features compared with points or edges. Using large image regions allows simple color classification with minimal training data and processing time, a particularly valid concern in earlier methods which were required to run on limited hardware. Additionally, region features are quite robust to occlusion and motion blur as they don't rely on fine scale features which are easily occluded or corrupted by scene movement. Methods of detecting pose with points or edges have advantages over region features when the image quality is good and there are few occlusions, which is why they have often been used to augment region features rather than replace them (Allan et al., 2014; Brox et al., 2010). They have particularly noticeable advantages over region based features as they provide much precise matching which, when correct matches are found, can provide higher detection accuracy than the coarse scale region features.

6.2.3. Prior Knowledge

When using fully data-driven learning techniques, appropriate object models are obtained directly from training samples. Thus, the method design is unaffected by object appearance conditions in the image and is instead dependent on the training strategy. Yet, many surgical tool detectors try to reduce the modelling complexity or deal with missing or challenging data with the integration of prior knowledge. However, challenges may occur when attempting to generalize over different sequences as shape or location constraints for one surgical instrument do not necessarily apply for others. Thus, a detection method designed for specific appearance conditions will not be robust enough to detect other tools or even the same one within another surgical context (e.g. different backgrounds). Nevertheless, for in-vivo surgical applications that are heavily reliant on surgical tool detection, the higher the performance the better the patient outcome even at the cost of reduced generalization. In that regard, adding as much prior knowledge as possible can be seen as a necessary cost.

6.2.4. Temporal Tracking

Temporal tracking has not been explored extensively in tool detection with most methods resorting to solving the problem with tracking-by-detection. This involves ignoring the temporal element of the problem instead treating each frame as a separate detection problem. Tracking by detection is useful when working in complex environments when the object being tracked may move in and out of view regularly which normally has to be handled explicitly in a temporal model with a reinitialization procedure. In the methods which do use temporal information, particle filters have proved more popular than Bayesian filters. This is most likely due to the ability of particle filters to reason about multiple hypotheses for the instrument state, which enables them to track effectively in cluttered environments. Kalman filters provide an easy to use, computationally cheap method of rigorously fusing temporal sensor data but are limited in that they can only provide accurate results for linear motion models with Gaussian uncertainties. However, for many applications in instrument detection this has been sufficient. Tracking-by-initialisation has also been explored and is particularly popular in methods which focus more strongly on the detection aspect of the solution. Although conceptually simple, tracking by initialisation is limited because it suffers from the drift associated with most tracking methods yet doesn't incorporate reasoning about a motion model or uncertainty which can provide a better solution.

6.2.5. Optimization Strategies

Processing data in real-time is crucial for integration into medical applications such as context-aware computer-assisted intervention (CACAI) systems. At the very least, a detection method should be able to process data at the same speed as the video acquisition sources to make a full use of available information. Obviously the main goal remains to obtain highly accurate results before being able to run fast. Unfortunately, comparison of surgical tool detection techniques based on processing speed is hard to achieve because of variation in image resolution and parameter search range. However, thanks to data-sets of reference and a standardized search space (e.g. scales per octave), detectors can be straightforwardly ranked and compared based on their speed performance. As a result, the frame per second processing capacity has been mentioned side-by-side with the Log-Average Miss-Rate value in the literature (Dollár et al., 2011).

Data processing speed-up can be achieved through either code optimization, parallel computing or the use of ad-hoc optimization strategies. Computer hardware specifications and parallel computing contribute the most to processing speed-up, especially when relying on an extensive use of the GPU. As such, every new hardware generation is accompanied by a huge speed boost. Regarding code optimization, investigating both CPU and GPU code is necessary, depending on physical hardware constraints. Some medical devices are currently equipped with a CPU only and can not integrate a GPU because of space restrictions. Waiting for high-end GPUs to enter the operating room might take some time, thus investing other speed-up solutions is unavoidable. In addition to space related issues, GPUs require high power-supply and multiple fans to reduce the heat from an intense use, which might be an issue within sterile environment where dust can not be expelled anywhere. As an alternative to hardware based improvements, ad-hoc optimization strategies can be used but the impact on detection accuracy should be assessed in order to find the best trade-off between speed and accuracy. For instance, cascade-based classifiers such as random forests can be speed-up by either limiting each tree depth, using some sort of soft-cascading (Benenson et al., 2012) or early stopping scheme (Sznitman et al., 2014). The most popular speed-up strategy is to perform down-sampling over input images or to use larger strides at runtime, for example processing every fourth line and column.

6.2.6. Detection and Classification

Currently, all methods of detection and pose estimation involve applying some known features of an object model to a target image. Typical approaches only focus on localizing one instrument type with its specific appearance model. When attempting to localize many different medical instruments, particular challenges might occur if the inter-class appearance variance is small and in such cases, false positive detections are common. Alternatively, it could be possible to find a representation of instruments which was sufficiently generic to allow most instruments to be localized accurately yet was specific enough to still localize accurately onto the target shape.
One approach of solving this is to perform recognition of the instrument type from its appearance before selecting the closest instrument model from a library such as has been proposed in (Speidel et al., 2009). Here, the authors perform basic recognition using an extracted contour of the tracked instrument. They learn principal components of the appearance model and use these as a minimal representation of each contour. Matching to a library is done by minimizing a distance metric.

6.3. Validation Methodologies, Protocols and Metrics

The purpose of any validation methodology is to quantify and rank detection methods performance in a realistic, unbiased and informative manner. To further ensure consistency and reproducibility in the results, using the exact same validation code is strongly encouraged, as opposed to individual re-implementations where key elements can be omitted or overlooked (Dollár et al., 2011). Currently, judging or comparing surgical tool detection approaches is hard because results are provided according to paper-specific validation methodologies and as such out of a reference context.

Validation typically focusses on 2D measurement due to the fact that single monocular input images are processed. Yet, some studies performed validation in the 3D space, but in such case not only stereoscopic input images are needed to obtain 3D tool detections but a real 3D ground-truth obtained via an automatic system is also necessary.

While quantitative results are most usually produced, some studies additionally report qualitative assessments. This is much more common in studies that include in-vivo data which does not have associated ground truth (Allan et al., 2015). Although this is potentially interesting in order to grasp some insights about detection success and failure modes, such results are heavily observer-biased and not at all reproducible. A proper methodology for comparison and ranking can only be performed through quantitative assessment.

Regarding the model validation strategy, a minority of studies have been using clear and explicit separation patterns between train and test image pools (e.g. k-fold cross-validation scheme) (Kumar et al., 2013a). Well described data-splitting scenarios can be used for validation of existing and pre-trained detectors, but also for creating new models (Dollár et al., 2011). Proper and even image distribution between train and test splits is mandatory to prevent learned surgical tool models from over-fitting the data. However, major drawbacks have been identified in previous studies where the beginning of an image sequence is used as part of train set and subsequent images of the same sequence are used as part of the test set. In addition, many times no clear explanation about image distributions is given and images from the training set may also be included in the test set. On a side note, online learning approaches do not require a clear model validation strategy since the tool model is being built and updated on-the-fly.

For computation, the choice of the metric, the reference, and the figure of merit are unsurprisingly all entangled and highly correlated to the specification phase. In case of a 3D validation, the Intersection Over Union criterion metric can not be used, mainly because 3D reference bounding boxes do not exist. Similarly, the number of consecutive frames metric can be used to assess tracking performance only.

Many times, the validation has focused on one tool landmark exclusively, such as the tool-tip (Allan et al., 2013) which can mislead about the total accuracy in estimation of the overall tool pose. Ideally all accessible tool pose parameters should be investigated, depending on tools degrees of freedom (Allan et al., 2015). Aside from global tool pose validation, it is worth mentioning interests in validating intermediate steps. For example, the pixel-wise classification (i.e. *semantic labelling*) is studied with per-class and per-pixel accuracy computation. Results provided are however somewhat flawed since the aesthetic quality is not captured and quantitatively equal results can be far from equal qualitatively. A computation more focused on visual quality would be probably more adequate.

Considering the temporal tracking aspect, the validation is solely performed through the computation of the number of consecutive frames successfully tracked. However, the added-value of the tracker itself has almost never been quantified. Comparing detector performance with and without the use of the tracking layer could be of interest. For object detection validation, a common set of parameters exists: the Intersection Over Union criterion metric, bounding boxes as reference, and recall, precision, and LAMR as figures of merit. Given the shape of surgical tools to detect and highly occurring in-plane rotations, bounding boxes usually do not fit well enough their appearances. In order to provide the utmost informative results, bounding boxes should be replaced by a tighter geometry (e.g. polygons) (Bouget et al., 2015). Additionally, depending on the nature of the tool to detect, and the relative size of its bounding geometry, it is necessary to modulate and thus trade lightly with the overlap threshold. For example in the literature, an overlap of 50% is required for pedestrians and cyclists detection validation, while for cars the overlap should be at least of 70% (Geiger et al., 2012).
Choosing a proper set of computation parameters should be driven by the final medical application in which the detection method will be integrated. In specific cases of tool positioning (e.g. needle insertion), only an accurate tool-tip location is mandatory, thus justifying the choice to perform exclusively an assessment over tool-tip location performance.

7. Conclusion

With the ever increasing use of MIS techniques there is a growing need for CAI systems in surgery. Automatic and accurate detection of surgical instruments within the coordinate system of surgical camera is critical and there are increasing efforts to develop image-based and marker-less tool detection approaches. In this paper, we have reviewed the state of the art in this field. We have discussed how computer vision techniques represent a highly promising approach of detecting, localizing and tracking instruments without the requirements of modifying the instrument design or interfering with the surgical work-flow. However, there are numerous outstanding technical challenges which must be addressed to enhance robustness in the presence of challenging conditions such as occlusion, blood, smoke, other instruments and the numerous other hazards which routinely occur within the field of view. Algorithms benchmarking and validation also needs to be enhanced with standardised data-sets and efforts are underway towards this end with for instance the Open-CAS website[13] trying to reference and collect in one place all existing surgical tool data-sets, or other dedicated websites [14] [15]. The community itself has been recently focusing on those two topics by initiating a call for data as part of the 2015 MICCAI conference. Addressing the aforementioned pressing matters should enable a faster growth in surgical tool detection performance and usage in novel computer assisted surgical systems. Only the existence of well-established and publicly available validation data-sets and methodologies can allow to properly measure progress. Evidently, a continuous evolution of both will be required to keep up with advances in the field. In addition, more sophisticated localization is required to fully model the articulation of many robotic instruments and techniques in human body pose and hand tracking (Prisacariu & Reid, 2011; Cashman & Fitzgibbon, 2013b; Cremers, 2006) have illustrated that this is possible using state-of-the-art computer vision methods.

8. Acknowledgements

David Bouget would like to acknowledge the financial support of Carl Zeiss Meditec AG. Max Allan would like to acknowledge the financial support of the Rabin Ezra foundation as well as the EPSRC funding for the DTP in Medical and Biomedical Imaging at UCL. Danail Stoyanov would like to acknowledge the financial support of a Royal Academy of Engineering/EPSRC Fellowship.

[13] http://opencas.webarchiv.kit.edu/?q=node/26

[14] http://www.surgicalvision.cs.ucl.ac.uk/benchmarking

[15] https://dbouget.bitbucket.org/2015_tmi_surgical_tool_detection

References

Abdel-Hakim, A. E., Farag, A. et al. (2006). Csift: A sift descriptor with color invariant characteristics. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (pp. 1978–1983). IEEE volume 2.

Ali, K., Fleuret, F., Hasler, D., & Fua, P. (2012). A real-time deformable detector. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*, 225–239.

Allan, M., Chang, P., Ourselin, S., Hawkes, D. J., Sridhar, A., Kelly, J., & Stoyanov, D. (2015). Image based surgical instrument pose estimation with multi-class labelling and optical flow. In N. W. Frangi, Hornegger (Ed.), *Medical Image Computing and Computer-Assisted Intervention MICCAI 2015* Lecture Notes in Computer Science. Springer International Publishing.

Allan, M., Ourselin, S., Thompson, S., Hawkes, D. J., Kelly, J., & Stoyanov, D. (2013). Toward detection and localization of instruments in minimally invasive surgery. *Biomedical Engineering, IEEE Transactions on*, *60*, 1050–1058.

Allan, M., Thompson, S., Clarkson, M. J., Ourselin, S., Hawkes, D. J., Kelly, J., & Stoyanov, D. (2014). 2d-3d pose tracking of rigid instruments in minimally invasive surgery. In *Information Processing in Computer-Assisted Interventions* (pp. 1–10). Springer.

Alsheakhali, M., Yigitsoy, M., Eslami, A., & Navab, N. (2015). Surgical tool detection and tracking in retinal microsurgery. In *SPIE Medical Imaging* (pp. 941511–941511). International Society for Optics and Photonics.

Ambai, M., & Yoshida, Y. (2011). Card: Compact and real-time descriptors. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (pp. 97–104). IEEE.

Babenko, B., Dollár, P., & Belongie, S. (2007). Task specific local region matching. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (pp. 1–8). IEEE.

Bardram, J. E., Doryab, A., Jensen, R. M., Lange, P. M., Nielsen, K. L., & Petersen, S. T. (2011). Phase recognition during surgical procedures using embedded and body-worn sensors. In *Pervasive Computing and Communications (PerCom), 2011 IEEE International Conference on* (pp. 45–53). IEEE.

Baumhauer, M., Feuerstein, M., Meinzer, H.-P., & Rassweiler, J. (2008). Navigation in endoscopic soft tissue surgery: perspectives and limitations. *Journal of Endourology*, *22*, 751–766.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (surf). *Computer vision and image understanding*, *110*, 346–359.

Benenson, R., Mathias, M., Timofte, R., & Van Gool, L. (2012). Pedestrian detection at 100 frames per second. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 2903–2910). IEEE.

Benenson, R., Mathias, M., Tuytelaars, T., & Van Gool, L. (2013). Seeking the strongest rigid detector. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on* (pp. 3666–3673).

Benenson, R., Omran, M., Hosang, J., & Schiele, B. (2014). Ten years of pedestrian detection, what have we learned? *arXiv preprint arXiv:1411.4304*, .

Bhattachayya, A. (1943). On a measure of divergence between two statistical population defined by their population distributions. *Bulletin Calcutta Mathematical Society*, *35*, 99–109.

Bibby, C., & Reid, I. (2008). Robust Real-Time visual tracking using Pixel-Wise posteriors. In *Proceedings of the 10th European Conference on Computer Vision* ECCV '08 (pp. 831–844).

Bouarfa, L., Akman, O., Schneider, A., Jonker, P. P., & Dankelman, J. (2012). In-vivo real-time tracking of surgical instruments in endoscopic video. *Minimally Invasive Therapy & Allied Technologies*, *21*, 129–134.

Bouget, D., Benenson, R., Omran, M., Riffaud, L., Schiele, B., & Jannin, P. (2015). Detecting surgical tools by modelling local appearance and global shape. *Medical Imaging, IEEE Transactions on*, *34*, 2603–2617.

Brox, T., Rosenhahn, B., Gall, J., & Cremers, D. (2010). Combined Region-and Motion-based 3D Tracking of Rigid and Articulated Objects. *IEEE Transactions on Pattern Analysis and Machine IntelligenceT*, *32*, 402–415. URL: `http://ieeexplore.ieee.org/xpls/abs\_all.jsp?arnumber=4775902\&amp;tag=1`.

Burschka, D., Corso, J. J., Dewan, M., Lau, W., Li, M., Lin, H., Marayong, P., Ramey, N., Hager, G. D., Hoffman, B. et al. (2005). Navigating inner space: 3-d assistance for minimally invasive surgery. *Robotics and Autonomous Systems*, *52*, 5–26.

Calonder, M., Lepetit, V., Strecha, C., & Fua, P. (2010). Brief: Binary robust independent elementary features. *Computer Vision–ECCV 2010*, (pp. 778–792).

Cano, A. M., Gayá, F., Lamata, P., Sánchez-González, P., & Gómez, E. J. (2008). Laparoscopic tool tracking method for augmented reality surgical applications. In *Biomedical Simulation* (pp. 191–196). Springer.

Casals, A., Amat, J., & Laporte, E. (1996). Automatic guidance of an assistant robot in laparoscopic surgery. In *Robotics and Automation, 1996. Proceedings., 1996 IEEE International Conference on* (pp. 895–900). IEEE volume 1.

Cashman, T., & Fitzgibbon, A. (2013a). What shape are dolphins? building 3D morphable models from 2D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 232–244. doi:`10.1109/TPAMI.2012.68`.

Cashman, T., & Fitzgibbon, A. (2013b). What shape are dolphins? building 3D morphable models from 2D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 232–244. doi:`10.1109/TPAMI.2012.68`.

Chmarra, M., Grimbergen, C., & Dankelman, J. (2007). Systems for tracking minimally invasive surgical instruments. *Minimally Invasive Therapy & Allied Technologies*, *16*, 328–340.

Christe, B., Cooney, E., Maggioli, G., Doty, D., Frye, R., & Short, J. (2008). Testing potential interference with rfid usage in the patient care environment. *Biomedical Instrumentation & Technology*, *42*, 479–484.

Cleary, K., & Nguyen, C. (2001). State of the art in surgical robotics: Clinical applications and technology challenges. *Computer Aided Surgery*, *6*, 312–328. URL: `http://dx.doi.org/10.1002/igs.10019`. doi:`10.1002/igs.10019`.

Cremers, D. (2006). Dynamical statistical shape priors for level set-based tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *28*, 1262–1273.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (pp. 886–893). IEEE volume 1.

Darzi, A., & Mackay, S. (2002). Recent advances in minimal access surgery. *BMJ: British Medical Journal*, *324*, 31.

Davis, B. (2000). A review of robotics in surgery. *Proceedings of the Institution of Mechanical Engineers. Part H, Journal of engineering in medicine*, *214*, 129–140. URL: `http://dx.doi.org/10.1243/09544119JEIM591`.

Dogangil, G., Davies, B., & Rodriguez, y. B. F. (2010). A review of medical robotics for minimally invasive soft tissue surgery. *Proceedings of the Institution of Mechanical Engineers. Part H, Journal of Engineering in Medicine*, *224*, 653–679. URL: `http://dx.doi.org/10.1243/09544119JEIM591`. doi:`10.1243/09544119JEIM591`.

Doignon, C., Graebling, P., & de Mathelin, M. (2005). Real-time segmentation of surgical instruments inside the abdominal cavity using a joint hue saturation color feature. *Real-Time Imaging*, *11*, 429–442.

Doignon, C., Nageotte, F., & De Mathelin, M. (2004). Detection of grey regions in color images: application to the segmentation of a surgical instrument in robotized laparoscopy. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on* (pp. 3394–3399). IEEE volume 4.

Doignon, C., Nageotte, F., & de Mathelin, M. (2007). Segmentation and guidance of multiple rigid objects for intra-operative endoscopic vision. In *Dynamical Vision* (pp. 314–327). Springer.

Dollar, P., Tu, Z., & Belongie, S. (2006). Supervised learning of edges and object boundaries. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (pp.

1964–1971). IEEE volume 2.

Dollár, P., Tu, Z., Perona, P., & Belongie, S. (2009). Integral channel features. In *Proceedings of the British Machine Vision Conference* (pp. 91.1–91.11). BMVA Press.

Dollár, P., Tu, Z., Tao, H., & Belongie, S. (2007). Feature mining for image classification. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (pp. 1–8). IEEE.

Dollár, P., Wojek, C., Schiele, B., & Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *TPAMI*, .

Duncan, R. G., & Froggatt, M. E. (2010). Fiber optic position and/or shape sensing based on rayleigh scatter.

Elfring, R., de la Fuente, M., & Radermacher, K. (2010). Assessment of optical localizer accuracy for computer aided surgery systems. *Computer Aided Surgery*, *15*, 1–12.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*, 861–874.

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1–8). IEEE.

Ferrari, V., Marin-Jimenez, M., & Zisserman, A. (2008). Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1–8). IEEE.

Flusser, J., & Suk, T. (1993). Pattern recognition by affine moment invariants. *Pattern recognition*, *26*, 167–174.

Fried, M. P., Kleefield, J., Gopal, H., Reardon, E., Ho, B. T., & Kuhn, F. A. (1997). Image-guided endoscopic surgery: Results of accuracy and performance in a multicenter clinical study using an electromagnetic tracking system. *The Laryngoscope*, *107*, 594–601.

Galleguillos, C., & Belongie, S. (2010). Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, *114*, 712–722.

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 3354–3361). IEEE.

Haase, S., Wasza, J., Kilgus, T., & Hornegger, J. (2013). Laparoscopic instrument localization using a 3-d time-of-flight/rgb endoscope. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on* (pp. 449–454). IEEE.

Haralick, R. M. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, *67*, 786–804.

Hartley, R. I., & Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. (2nd ed.). Cambridge University Press, ISBN: 0521540518.

Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., & Lepetit, V. (2011). Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (pp. 858–865). IEEE.

Houliston, B., Parry, D., Webster, C. S., & Merry, A. F. (2009). Interference with the operation of medical devices resulting from the use of radio frequency identification technology. *NZ Med J*, *122*, 9–16.

Hu, Y., Ahmed, H., Allen, C., Pends, D., Sahu, M., Emberton, M., Hawkes, D., & Barratt, D. (2009). Mr to ultrasound image registration for guiding prostate biopsy and interventions. In G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, & C. Taylor (Eds.), *Medical Image Computing and Computer-Assisted Intervention MICCAI 2009* (pp. 787–794). Springer Berlin Heidelberg volume 5761 of *Lecture Notes in Computer Science*. URL: http://dx.doi.org/10.1007/978-3-642-04268-3_97. doi:10.1007/978-3-642-04268-3_97.

Hue, C., Le Cadre, J.-P., & Pérez, P. (2000). *Tracking Multiple Objects with Particle Filtering*. Research Report RR-4033 INRIA. URL: http://hal.inria.fr/inria-00072605.

Isard, M., & Blake, A. (1998). Condensationconditional density propagation for visual tracking. *International journal of computer vision*, *29*, 5–28.

Jannin, P., Grova, C., & Maurer Jr, C. R. (2006). Model for defining and reporting reference-based validation protocols in medical image processing. *International Journal of Computer Assisted Radiology and Surgery*, *1*, 63–73.

Jannin, P., & Korb, W. (2008). Assessment of image-guided interventions. In *Image-Guided Interventions* (pp. 531–549). Springer.

Ko, S.-Y., Kim, J., Kwon, D.-S., & Lee, W.-J. (2005). Intelligent interaction between surgeon and laparoscopic assistant robot system. In *Robot and Human Interactive Communication, IEEE International Workshop on* (pp. 60–65). IEEE.

Kohn, L. T., Corrigan, J. M., Donaldson, M. S. et al. (2000). *To Err Is Human:: Building a Safer Health System* volume 627. National Academies Press.

Krähenbühl, P., & Koltun, V. (2012). Efficient inference in fully connected crfs with gaussian edge potentials. *CoRR*, *abs/1210.5644*. URL: http://arxiv.org/abs/1210.5644.

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

Krupa, A., Gangloff, J., Doignon, C., de Mathelin, M. F., Morel, G., Leroy, J., Soler, L., & Marescaux, J. (2003). Autonomous 3-d positioning of surgical instruments in robotized laparoscopic surgery using visual servoing. *Robotics and Automation, IEEE Transactions on*, *19*, 842–853.

Kumar, S., Narayanan, M., Misra, S., Garimella, S., Singhal, P., Corso, J., & Krovi, V. (2013a). Video-based framework for safer and smarter computer aided surgery. In *Hamlyn Symposium on Medical Robotics* (pp. 107–108).

Kumar, S., Narayanan, M. S., Singhal, P., Corso, J. J., & Krovi, V. (2013b). Product of tracking experts for visual tracking of surgical tools. In *Automation Science and Engineering (CASE), 2013 IEEE International Conference on* (pp. 480–485). IEEE.

Lahanas, V., Loukas, C., & Georgiou, E. (2015). A simple sensor calibration technique for estimating the 3d pose of endoscopic instruments. *Surgical Endoscopy*, (pp. 1–7).

Laptev, I. (2006). Improvements of object detection using boosted histograms. In *BMVC* (pp. 949–958). volume 6.

Lee, C., Wang, Y.-F., Uecker, D., & Wang, Y. (1994). Image analysis for automated tracking in robot-assisted endoscopic surgery. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision &amp; Image Processing., Proceedings of the 12th IAPR International Conference on* (pp. 88–92). IEEE volume 1.

Lepetit, V., Moreno-Noguer, F., & Fua, P. (2009). Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, *81*, 155–166.

Li, Y., Chen, C., Huang, X., & Huang, J. (2014). Instrument tracking via online learning in retinal microsurgery. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014* (pp. 464–471). Springer.

Lo, B. P., Darzi, A., & Yang, G.-Z. (2003). Episode classification for the analysis of tissue/instrument interaction with multiple visual cues. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003* (pp. 230–237). Springer.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on* (pp. 1150–1157). Ieee volume 2.

Maier-Hein, L., Mersmann, S., Kondermann, D., Bodenstedt, S., Sanchez, A., Stock, C., Kenngott, H. G., Eisenmann, M., & Speidel, S. (2014). Can masses of non-experts train highly accurate image classifiers? In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014* (pp. 438–445). Springer.

Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R. et al. (1999). Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop* (pp. 249–252).

Malik, J., Belongie, S., Leung, T., & Shi, J. (2001). Contour and texture analysis for image segmentation. *International journal of computer vision*, *43*, 7–27.

Mathias, M., Benenson, R., Timofte, R., & Gool, L. V. (2013). Handling occlusions with franken-classifiers. In *Computer Vision (ICCV), 2013 IEEE International Conference on* (pp. 1505–1512). IEEE.

McKenna, S., Charif, H. N., & Frank, T. (2005). Towards video understanding of laparoscopic surgery: Instrument tracking. In *Proc. of Image and Vision Computing New Zealand (IVCNZ)*.

McLachlan, G. (2004). *Discriminant analysis and statistical pattern recognition* volume 544. John Wiley & Sons.

Van der Meijden, O., & Schijven, M. (2009). The value of haptic feedback in conventional and robot-assisted minimal invasive surgery and virtual reality training: a current review. *Surgical endoscopy*, *23*, 1180–1190.

Mirota, D. J., Ishii, M., & Hager, G. D. (2011). Vision-based navigation in image-guided interventions. *Annual review of biomedical engineering*, *13*, 297–319.

Miyawaki, F., Tsunoi, T., Namiki, H., Yaginuma, T., Yoshimitsu, K., Hashimoto, D., & Fukui, Y. (2009). Development of automatic acquisition system of surgical-instrument informantion in endoscopic and laparoscopic surgey. In *Industrial Electronics and Applications, 2009. ICIEA 2009. 4th IEEE Conference on* (pp. 3058–3063). IEEE.

Moreno-Noguer, F., Lepetit, V., & Fua, P. (2007). Accurate non-iterative o (n) solution to the pnp problem. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (pp. 1–8). IEEE.

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, *24*, 971–987.

Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, *11*, 23–27.

Papageorgiou, C. P., Oren, M., & Poggio, T. (1998). A general framework for object detection. In *Computer vision, 1998. sixth international conference on* (pp. 555–562). IEEE.

Parlak, S., Marsic, I., & Burd, R. S. (2011). Activity recognition for emergency care using rfid. In *Proceedings of the 6th International Conference on Body Area Networks* (pp. 40–46). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Pezzementi, Z., Voros, S., & Hager, G. D. (2009). Articulated object tracking by rendering consistent appearance parts. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on* (pp. 3940–3947). IEEE.

Prince, S. J. (2012). *Computer vision: models, learning, and inference*. Cambridge University Press.

Prisacariu, V. A., & Reid, I. (2011). Nonlinear shape manifolds as shape priors in level set segmentation and tracking. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 2185–2192). IEEE.

Prisco, G. (2010). Fiber optic shape sensor.

Reiter, A., & Allen, P. K. (2010). An online learning approach to in-vivo tracking using synergistic features. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on* (pp. 3441–3446). IEEE.

Reiter, A., Allen, P. K., & Zhao (2012a). Marker-less articulated surgical tool detection. In *Proc. Computer Assisted Radiology and Surgery* (pp. 175–176). volume 7.

Reiter, A., Allen, P. K., & Zhao, T. (2012b). Articulated surgical tool detection using virtually-rendered templates. In *Computer Assisted Radiology and Surgery (CARS)*.

Reiter, A., Allen, P. K., & Zhao, T. (2012c). Feature classification for tracking articulated surgical tools. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012* (pp. 592–600). Springer.

Reiter, A., Allen, P. K., & Zhao, T. (2012d). Learning features on robotic surgical tools. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on* (pp. 38–43). IEEE.

Richa, R., Balicki, M., Meisner, E., Sznitman, R., Taylor, R., & Hager, G. (2011a). Visual tracking of surgical tools for proximity detection in retinal surgery. In *Information Processing in Computer-Assisted Interventions* (pp. 55–66). Springer.

Richa, R., Sznitman, R., Taylor, R., & Hager, G. (2011b). Visual tracking using the sum of conditional variance. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference On* (pp. 2953–2958). IEEE.

Rieke, N., Tan, D., Alsheakhali, M., Tombari, F., di San Filippo, C., Belagiannis, V., Eslami, A., & Navab, N. (2015). Surgical tool tracking and pose estimation in retinal microsurgery. In N. Navab, J. Hornegger, W. Wells, & A. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention MICCAI 2015* (pp. 266–273). Springer International Publishing volume 9349 of *Lecture Notes in Computer Science*. URL: `http://dx.doi.org/10.1007/978-3-319-24553-9_33`. doi:10.1007/978-3-319-24553-9_33.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., & Moore, R. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, *56*, 116–124.

Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2009). Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int. Journal of Computer Vision (IJCV)*, . URL: `http://research.microsoft.com/apps/pubs/default.aspx?id=117885`.

Speidel, S., Benzko, J., Krappe, S., Sudra, G., Azad, P., Müller-Stich, B. P., Gutt, C., & Dillmann, R. (2009). Automatic classification of minimally invasive instruments based on endoscopic image sequences. In *SPIE Medical Imaging* (pp. 72610A–72610A). International Society for Optics and Photonics.

Speidel, S., Delles, M., Gutt, C., & Dillmann, R. (2006). Tracking of instruments in minimally invasive surgery for surgical skill analysis. In *Medical Imaging and Augmented Reality* (pp. 148–155). Springer.

Speidel, S., Krappe, S., Röhl, S., Bodenstedt, S., Müller-Stich, B., & Dillmann, R. (2013). Robust feature tracking for endoscopic pose estimation and structure recovery. In *SPIE Medical Imaging* (pp. 867102–867102). International Society for Optics and Photonics.

Speidel, S., Kuhn, E., Bodenstedt, S., Röhl, S., Kenngott, H., Müller-Stich, B., & Dillmann, R. (2014). Visual tracking of da vinci instruments for laparoscopic surgery. In *SPIE Medical Imaging* (pp. 903608–903608). International Society for Optics and Photonics.

Speidel, S., Sudra, G., Senemaud, J., Drentschew, M., Müller-Stich, B. P., Gutt, C., & Dillmann, R. (2008). Recognition of risk situations based on endoscopic instrument tracking and knowledge based situation modeling. In *Medical Imaging* (pp. 69180X–69180X). International Society for Optics and Photonics.

Stoyanov, D. (2012). Surgical vision. *Annals of biomedical engineering*, *40*, 332–345.

Sznitman, R., Ali, K., Richa, R., Taylor, R. H., Hager, G. D., & Fua, P. (2012). Data-driven visual tracking in retinal microsurgery. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012* (pp. 568–575). Springer.

Sznitman, R., Basu, A., Richa, R., Handa, J., Gehlbach, P., Taylor, R., Jedynak, B., & Hager, G. (2011). Unified detection and tracking in retinal microsurgery. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*, (pp. 1–8).

Sznitman, R., Becker, C. J., & Fua, P. (2014). Fast part-based classification for instrument detection in minimally invasive surgery. In *Medical Image Computing and Computer Assisted Intervention*. Springer.

Sznitman, R., Richa, R., Taylor, R. H., Jedynak, B., & Hager, G. D. (2013). Unified detection and tracking of instruments during retinal microsurgery. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *35*, 1263–1273.

Tonet, O., Thoranaghatte, R. U., Megali, G., & Dario, P. (2007). Tracking endoscopic instruments without a localizer: A shape-analysis-based approach. *Computer Aided Surgery*, *12*, 35–42.

Tu, Z. (2005). Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* (pp. 1589–1596). IEEE volume 2.

Tu, Z., Narr, K. L., Dollár, P., Dinov, I., Thompson, P. M., & Toga, A. W. (2008). Brain anatomical structure segmentation by hybrid discriminative/generative models. *Medical Imaging, IEEE Transactions on*, *27*, 495–508.

Van De Weijer, J., Schmid, C., Verbeek, J., & Larlus, D. (2009).

Learning color names for real-world applications. *Image Processing, IEEE Transactions on*, *18*, 1512–1523.

Van Der Togt, R., van Lieshout, E. J., Hensbroek, R., Beinat, E., Binnekade, J., & Bakker, P. (2008). Electromagnetic interference from radio frequency identification inducing potentially hazardous incidents in critical care medical equipment. *Jama*, *299*, 2884–2890.

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, *44*, 330–349.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, *57*, 137–154.

Voros, S., Long, J.-A., & Cinquin, P. (2007). Automatic detection of instruments in laparoscopic images: A first step towards high-level command of robotic endoscopic holders. *The International Journal of Robotics Research*, *26*, 1173–1190.

Weese, J., Penney, G. P., Desmedt, P., Buzug, T. M., Hill, D. L., & Hawkes, D. J. (1997). Voxel-based 2-d/3-d registration of fluoroscopy images and ct scans for image-guided surgery. *Information Technology in Biomedicine, IEEE Transactions on*, *1*, 284–293.

Wei, G.-Q., Arbter, K., & Hirzinger, G. (1997). Real-time visual servoing for laparoscopic surgery. controlling robot motion with color image segmentation. *Engineering in Medicine and Biology Magazine, IEEE*, *16*, 40–45.

Wolf, R., Duchateau, J., Cinquin, P., & Voros, S. (2011). 3d tracking of laparoscopic instruments using statistical and geometric modeling. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011* (pp. 203–210). Springer.

Yang, M., Kpalma, K., & Ronsin, J. (2008). A survey of shape feature extraction techniques. *Pattern recognition*, (pp. 43–90).

Zhang, X., & Payandeh, S. (2002). Application of visual tracking for robot-assisted laparoscopic surgery. *Journal of Robotic systems*, *19*, 315–328.

Zhou, J., & Payandeh, S. (2014). Visual tracking of laparoscopic instruments. *Journal of Automation and Control Engineering Vol*, *2*.