Synchronising a stereoscopic surgical video stream using specular reflection

Kilian Chandelon, Adrien Bartoli

Synchronising a Stereoscopic Surgical Video Stream using Specular Reflection

Kilian Chandelon[1,2*] and Adrien Bartoli[3,1,2]

[1]EnCoV, Institut Pascal, UMR6602 CNRS, UCA, Clermont-Ferrand University Hospital, France.
[2]SURGAR - Surgical Augmented Reality, Clermont-Ferrand, France.
[3]Department of Clinical Research and Innovation, Clermont-Ferrand University Hospital, France.

*Corresponding author(s). E-mail(s): kilian.chandelon@gmail.com;

Abstract

Purpose. A stereoscopic surgical video stream consists of left-right image pairs provided by a stereo endoscope. While the surgical display shows these image pairs synchronised, most capture cards cause de-synchronisation. This means that the paired left and right images may not correspond once used in downstream tasks such as stereo depth computation. The stereo synchronisation problem is to recover the corresponding left-right images. This is particularly challenging in the surgical setting, owing to the moist tissues, rapid camera motion, quasi-staticity and real-time processing requirement. Existing methods exploit image cues from the diffuse reflection component and are defeated by the above challenges.
Methods. We propose to exploit the specular reflection. Specifically, we propose a powerful left-right comparison score (LRCS) using the specular highlights commonly occurring on moist tissues. We detect the highlights using a neural network, characterise them with invariant descriptors, match them, and use the number of matches to form the proposed LRCS. We perform evaluation against 147 existing LRCS in 44 challenging robotic partial nephrectomy and robotic assisted hepatic resection video sequences with simulated and real de-synchronisation.
Results. The proposed LRCS outperforms, with an average and maximum offsets of 0.055 and 1 frames and 94.1±3.6% successfully synchronised frames. In contrast, the best existing LRCS achieves an average and maximum offsets of 0.3 and 3 frames and 81.2±6.4% successfully synchronised frames.
Conclusion. The use of specular reflection brings a tremendous boost to the real-time surgical stereo synchronisation problem.

Keywords: stereo, synchronisation, specularity, endoscopy, left-right comparison score

1 Introduction

Robot-assisted surgery is characterised by reduced invasiveness, improved patient recovery and superior clinical outcome compared to open surgery [1]. Surgical robots also improve the surgeon's perception of the operating site: they implement stereoscopic cameras which address the lack of depth perception associated with monocular laparoscopic cameras. However, accurately localising intraparenchymal anatomical structures for preservation or removal remains an issue. Recent advances in computer-assisted surgery (CAS) mitigate this issue by implementing augmented reality to overlay the video stream with the internal anatomical structures [2, 3]. CAS requires real-time processing of the surgical video stream, typically running on a powerful computer equipped with a video capture card and a GPU. The stereo video stream is composed of stereo images, which are left-right image pairs, captured at a frequency typically comprised within 20-60 Hz. The video processing unit triggers the image capture and obtains synchronous left-right images using a synchronisation signal and a time-stamping protocol. A synchronous video stream is then displayed in the surgical console and exposed for use by third-party systems, including CAS systems. At this point, maintaining synchrony can no longer be guaranteed because the link to the internal time-stamping system is severed. De-synchronisation then frequently arises from the CAS system's processing chain, both owing to the hardware components and the multithreaded programming of the capture card, as shown in figure 1. De-synchronisation has dramatic consequences on CAS systems and more generally on most downstream tasks, such as stereo depth computation. Consequently, a synchronisation processing must be used prior to image processing in the CAS system.

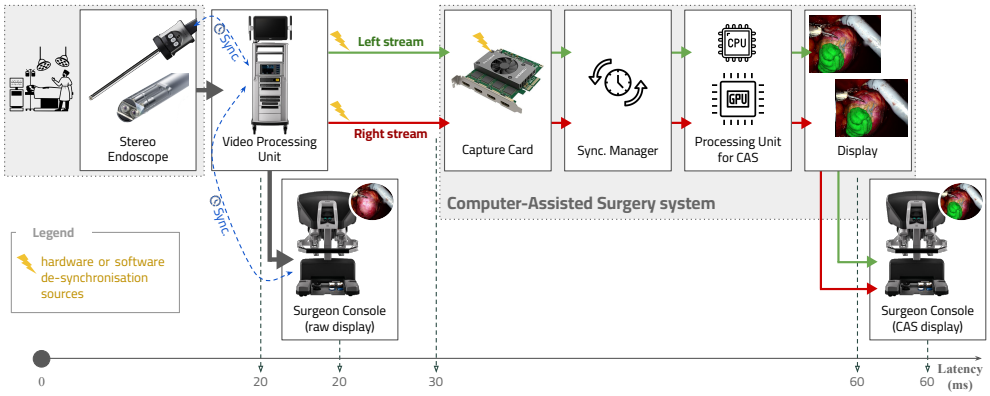


Fig. 1 Surgical data flow from a stereoscopic endoscope to the surgical console and subsequent integration into a CAS system. Introducing hardware or software components into the processing chain may result in de-synchronisation, significantly preventing the integration of surgical technology.

Surgical stereo synchronisation is thus a core challenge in the development of CAS systems. Synchronisation methods can be broadly classified as hardware-based or software-based. Hardware-based synchronisation relies on introducing binary-state

light sources in the field of view, on GPS positioning [4] or on oscillating objects [5] as external references. In the surgical setting, the deployment of new equipment is however restricted by compliance to medical device regulations and aseptic guidelines. The modification of existing material is prohibited and the use of existing data sources is favoured. This precludes the use of hardware-based synchronisation. Software-based synchronisation relies on the image contents such as brightness-change [6] or image descriptors [7] and is thus well-adapted to the surgical setting. However, surgical stereo synchronisation presents specific challenges related to the video contents, including (C1) quasi-static video, (C2) fast camera motion, and (C3) complex light reflection produced by surgical tools and moist surfaces. The challenges are also related to two requirements: real-time processing, which increases the complexity by demanding fast computations so as to not affect downstream tasks by adding latency, and autonomy, which prevents the method to rely on user intervention.

Technically, stereo synchronisation uses a high-level *synchronisation manager*. It receives the de-synchronised left and right channels, performs left-right image comparisons, selects the best matching image pairs, and outputs synchronised channels. It searches for the best corresponding right image for any given left image and vice versa. This relies on an optimisation problem minimising the overall temporal difference whilst maintaining the temporal consistency by preserving the image order, and tagging the images which do not have a match in the other channel. The core requirement, common to all methods, is the ability to quantify the temporal difference between a left and a right images. This quantification is achieved by a *Left-Right Comparison Score* (LRCS). While the synchronisation manager is fairly generic to the use-case, the LRCS, based on the image contents, must be specifically adapted. Existing LRCS are unadapted to the surgical setting. They typically use keypoint correspondences or global image descriptors. They thus rely on the diffuse reflection. The observed pixel intensity is the result of the cold polychromatic light source on board the endoscope reflected by the tissue surface. This reflection has two modes, which are combined to produce the final observed intensity. The diffuse reflection, which is fairly independent of the viewpoint, is the most common and tends to dominate in non-medical images. It allows surface colours to be perceived. The specular reflection, which strongly depends on the viewpoint, is highly typical of metal and moist surfaces, hence significant in surgical images. It is generally associated with sensor saturation. Ignoring the specular component in synchronisation has two negative consequences: first, it perturbs the diffuse measurements, and second, it wastes a good deal of potentially useful cues.

We propose to use the specular component to devise an LRCS particularly adapted to the surgical setting. Concretely, the specular component manifests itself primarily as specular highlights, appearing as white blobs in the images. Our specular LRCS hinges on the specular highlights, using a complete proposed pipeline that includes their detection, description and matching. We have validated our approach by performing experiments against 147 existing LRCS in 44 challenging robotic partial nephrectomy and robotic assisted hepatic resection videos with simulated and real de-synchronisation. This confirms our hypothesis that coupling the strong sensitivity to micro-movements of specular highlights with traditional diffuse image analysis provides an effective solution to the surgical stereo synchronisation problem.

2 Previous Work

We review image-based software synchronisation methods without restricting to surgery, with a special focus on the image cues used in the LRCS and their applicability in surgery with respect to the above-defined three challenges (C1), (C2) and (C3). The methods fall into two categories, depending of whether they use a single image or a sequence of images. In other words, whether they do not or do use temporal information. The single-image category includes global image descriptors such as self-similarity matrices (SSM) [8], frequency analysis and phase correlation [9], fingerprinting [10] and histogram [11] correlation. This category also includes local image descriptors by means of keypoints such as SIFT [7, 12, 13]. The image-sequence category includes brightness-change [6, 14], motion [15–17], activity [18] and compression bitrate profile [19] analysis. These methods are all based on the diffuse reflection, hence defeated by challenges (C1) and (C3). For instance, keypoints are highly sensitive to the specular highlights. Most single-image methods are defeated by challenge (C2), in contrast to most image-sequence methods, with the exception of brightness-change analysis, which carries limited significance due to the controlled and consistent artificial lighting within the cavity. Bitrate profile analysis has a significant computational load, which constitutes a challenge for the real-time applicability. We propose the first LRCS using both diffuse and specular reflection applicable to surgical stereo synchronisation. In contrast to existing work, it handles the three challenges and the resulting method easily integrates in third-party devices in the surgical workflow.

3 Materials and Methods

We first describe some general points to better understand the origin of desynchronisation. We then describe the synchronisation manager and finally the proposed specular LRCS and combined diffuse-specular LRCS.

3.1 General Points

In robotic endoscopic surgery, a stereoscopic acquisition and processing chain allows the surgeon to observe the surgical site binocularly. This view is only possible by using a specific hardware setup including a stereo surgical endoscope. Once the endoscope is inserted in the cavity via a trocar, the tissue surfaces are illuminated by a cold polychromatic light source on board the 3D endoscope. The light rays are projected onto the biological surfaces. Part of the light emitted is absorbed or transmitted and the rest is reflected. The physics of light-matter interaction is highly complex. It depends, among other things, on the nature of the materials, the characteristic properties of the interfaces of the material in which the electromagnetic radiation is propagated and the wavelength of the incident rays. However, we are only concerned with the reflected light, as this accounts for the vast majority of the light captured by the imaging system. In more detail, there are two combined modes of reflection: diffuse reflection, which once captured enables surface colours to be perceived, and specular reflection, which creates specular spots on the image that are generally associated with sensor saturation. The light captured by the sensor, an ordered arrangement of photodiodes,

is then converted into electrical intensity via the photoelectric effect. The analogue signal is then amplified and digitised before being fed into a digital video processing chain. In a stereo system, two sensors simultaneously capture images of the surgical site at a frequency generally comprised between 20 Hz and 60 Hz. The video processing unit of the robotic system is responsible for triggering the image capture, processing the signals until a pair of synchronous left-right images is obtained. Synchrony is generally achieved by using a synchronisation signal and a time-stamping protocol. A synchronous surgical video stream is then displayed to the surgeon in the surgical console and exposed for use by third-party systems. At this point, maintaining synchrony can no longer be guaranteed because the link to the internal time-stamping system is severed. In addition, downstream processing of the robotic system's video output, such as recording or displaying augmented reality, can in turn have negative effects on the synchrony of the left and right channels.

In CAS systems using stereo, a reliable left-right synchronisation is important, for example to perform stereo 3D reconstruction. In other words, if the left and right channels are out of sync, a synchronisation treatment must be applied before the images are processed by the CAS system.

3.2 Synchronisation Manager

The synchronisation manager we use is shown in figure 2. It implements the synchronisation mechanism which, from an input of de-synchronised image pairs $[F_{L,t}, F_{R,t}]$, where t is the time index, outputs synchronous image pairs $[\hat{F}_{L,t}, \hat{F}_{R,t}]$ at a frequency f_{out} equivalent to the input frequency f_{in}, while introducing a latency required to be lower than the period $T = 1/f_{\text{in}}$ between two consecutive images in the video stream.

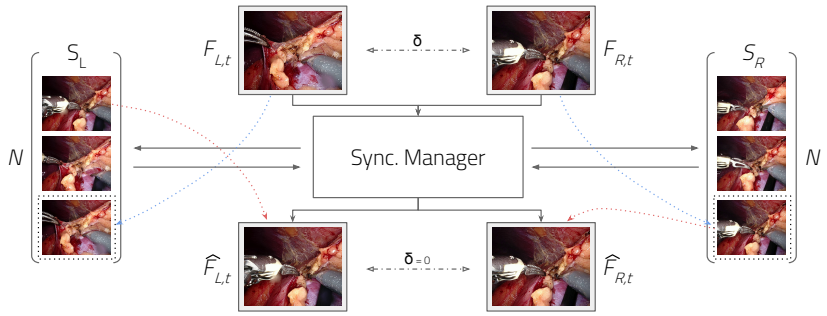


Fig. 2 Overview of the synchronisation manager, connecting a de-synchronised stereo input $[F_{L,t}, F_{R,t}]$ to a synchronised stereo output $[\hat{F}_{L,t}, \hat{F}_{R,t}]$ by using buffers S_L and S_R and the LRCS δ.

The synchronisation manager uses two buffer memories S_L and S_R of size N respectively storing the past N left and right analysed frames. To ensure that system speed is maintained and adverse effects are prevented, it is important to maintain N at a reasonable value, typically ranging between $f_{\text{in}}/6$ and 120. The synchronisation manager evaluates the temporal difference δ between the input pair frames and seeks to minimise it. To achieve this, three successive steps are required: 1) to detect and describe

the left and right image cues, 2) to carry out cues matching tests by evaluating the LRCS and 3) to choose the pair of frames that minimises the temporal distance δ.

The current left $F_{L,t}$ and right $F_{R,t}$ input images are first concatenated with the two buffer memories S_L and S_R respectively. Then $F_{L,t}$ is compared to all images in S_R and $F_{R,t}$ to all images in S_L using multithreading. Two resulting vectors C_L and C_R are created containing the LRCS. The left and right frames forming the synchronous pair are eventually selected by taking the frame pair which minimises the C_L and C_R and which also meets the non-return time condition. This is achieved by purging the posterior images from the memory buffers each time a synchronous pair is generated. Temporal smoothing is generally added to the process. We use the Exponentially Weighted Moving Average (EWMA). Pseudo-code is given in algorithm 1 for an example descriptor.

Algorithm 1 Sync Manager Algorithm

function SYNCMANAGER
  for each $\langle F_{L,t}, F_{R,t} \rangle$ in parallel **do**
    $H_L, H_R \leftarrow$ DETECT_HIGHLIGHT_BLOBS$(F_{L,t}, F_{R,t})$
    $d1_L, d1_R \leftarrow$ COMPUTE_SPEC_DAISY(H_L, H_R)
    $d2_L, d2_R \leftarrow$ COMPUTE_SIFT(H_L, H_R)
    $D_L \leftarrow$ CONCAT$(d1_L, d2_L)$
    $D_R \leftarrow$ CONCAT$(d1_R, d2_R)$
    if size of $S_L > 0$ **then**
      $C_L \leftarrow$ COMPUTE_LRCS(D_L, S_L)
      $C_R \leftarrow$ COMPUTE_LRCS(D_R, S_R)
      $I, S \leftarrow$ SELECTION(MAX(C_L), MAX(C_R))
      $I \leftarrow$ EWMA$(I, \text{prev}I)$
      Append F_L to S_L
      Append F_R to S_R
      Append I to prev$I$
      if $S =$ 'left' **then**
        $\hat{F}_{L,t} \leftarrow S_L(I)$
      else
        $\hat{F}_{L,t} \leftarrow F_{L,t}$
      end if
      if $S =$ 'right' **then**
        $\hat{F}_{R,t} \leftarrow S_R(I)$
      else
        $\hat{F}_{R,t} \leftarrow F_{R,t}$
      end if
    end if
  end for
  return $\langle \hat{F}_{L,t}, \hat{F}_{R,t} \rangle$
end function

3.3 The Specular LRCS

For pure specular reflection analysis, we propose to detect the specular highlight spots and describe them one by one as uniquely as possible so that they can be matched robustly by comparing their descriptors. We use the two buffer memories S_L and S_R to hold the images and for each, a list of highlights and associated descriptors.

In technical terms, the automatic detection of specular highlight spots is performed on each left and right image using a fully convolutional network running in parallel and postprocessing. For an input image, the network produces a binary mask M which specifically finds highlight pixels. We propose a network architecture composed of 18 convolution layers with 64 filters, a kernel size of (3, 3), strides of (1, 1) and an orthogonal kernel initialiser, as shown in figure 3. It uses ReLU activations and batch normalisation with a momentum of 0.1 and a minimal value of 0.0001, and sigmoid activation in the last layer. Overall, the network has 594 240 trainable and 2 048 non-trainable parameters. For training we created a dataset of 1000 images extracted from 10 robotic assisted partial nephrectomy (RAPN) procedures which we manually annotated using adaptive thresholding. We split the dataset in 800 training and 200 validation images. We used data augmentation with standard image transformations, namely horizontal and vertical flipping and realistic brightness adjustment within range $[0.3, 0.3]$. We trained with Adam optimiser for 20 epochs with a batch size of 32, a learning rate of 0.001 and a binary cross-entropy loss function.

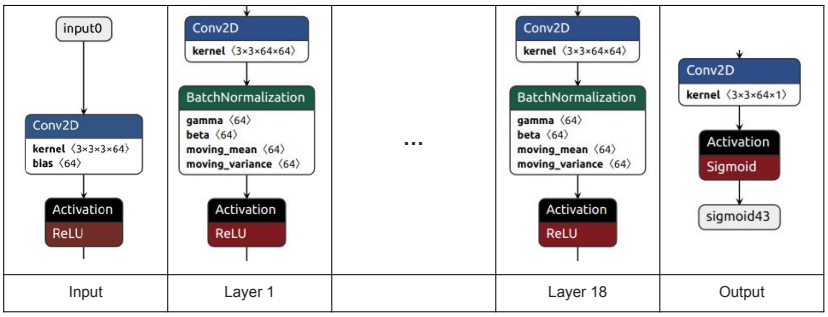


Fig. 3 The proposed neural network architecture for specular highlight segmentation.

We then run an isolation step, by instantiating a highlight spot for each connected component in M. In a parallel iterative loop, each spot is characterised by 5 components: 3 geometric properties, namely the position of its centroid, its area and its first moment, and 2 intensity properties analysed on a square patch of 10 pixels side centred on the centroid of the specular spot, namely its 10 bins colour histogram and the pure DAISY descriptor [20]. Then a cross-checked brute force matching algorithm from OpenCV is used to compare the spot descriptors. The LRCS is given by the number of matches passing the Lowe ratio test with $d = 0.8$.

3.4 The Diffuse-Specular LRCS

In order to exploit both the specular and the diffuse reflection components to improve the stability and accuracy of synchronisation, we propose a combined diffuse-specular LRCS. We achieve this by adding, for each left-right image pair to compare, the proposed specular LRCS and an existing diffuse LRCS. For the diffuse LRCS, we have chosen the number of SIFT matches [21], which combines reasonable invariance and fast computation.

4 Experimental Results

4.1 Dataset

The evaluation dataset includes 44 stereoscopic sequences collected during Robotic-Assisted Partial Nephrectomy (RAPN) and Robotic Assisted Hepatic Resection (RAHR) performed using the Da Vinci Xi surgical robot. It comprises 38 RAPN and 6 RAHR videos distributed in 26 short sequences with 21 frames and 18 extended sequences with 200 frames. All sequences were captured using RGB colour endoscopic imaging, except for 4 short RAPN sequences captured using indocyanine green (ICG) fluorescence. These sequences vividly describe surgical scenarios encountered during the successive phases of RAPN or RAHR, each characterised by distinct camera movements and instrument manipulations, providing a versatile foundation for analysis.

Synchronisation relies on a customised stereoscopic video recorder with a fixed frame rate and trigger system to grab and retrieve frames quasi-synchronously. A meticulous frame-by-frame visual analysis is then performed, using events like electrosurgical arcs as temporal cues for precise temporal alignment. This rigorous approach guarantees accurate synchronisation for robust algorithm development and evaluation.

4.2 Simulating De-synchronisation

We have developed a de-synchronisation simulator that transmutes a synchronous stereo sequence into an asynchronous semi-synthetic one, with exact control of the de-synchronisation parameters. This encompasses the simulation of Variable Frame Rate (VFR), stochastic frame dropping and initial offset. These simulated conditions provide a meticulously controlled experimental framework for algorithmic evaluation.

In greater elaboration, the de-synchronisation simulator addresses frame indices in relation to three parameters: `left_offset`, `right_offset` and `vfr_range`, which are specified for each simulation run by the user. Typical values used to emulate existing systems that capture and process a 60 Hz stereoscopic stream encompass a value ranging from 0 to 120 frames for offsets[1], and a range of [-1, 1] frames for VFR. Throughout the simulation process, `left_offset` and `right_offset` are initially applied to their respective left and right channels by discarding the initial

[1] On-board surgical recording systems such as Medicapture's MVR Pro, that use multithreading, combine two desynchronisation sources: 1) an initial offset, forming a constant shift, and 2) a variable offset. The initial offset is primarily caused by delays in thread start-up while the variable offset is due to threads being unsynchronised in software and to limited hardware capacity.

frames according to the provided values. Subsequently, if VFR simulation is enabled, subsampling of the original left and right synchronised sequences occurs to enable the simulation of frame dropping. The subsampling factor is determined as $\nu = 2\max(|\texttt{vfr\_range}[0]|, |\texttt{vfr\_range}[1]|) + 1$. In other words, this refers to the step size between the indices of two consecutive frames within the initial synchronous sequence, allowing VFR to be simulated within the given `vfr_range` without duplication or overlapping.

The de-synchronised frame indices are derived by introducing stochastic noise defined within `vfr_range` to the original frame indices. Ultimately, the left and right frame channels undergo processing to ensure equal length, after which they are exported as image files for ease of use.

4.3 Synchronisation Performance

The investigation of synchronisation performance is conducted through the presentation of both quantitative and qualitative outcomes, encompassing response curves and a comparative analysis between synchronisation methods.

4.3.1 Evaluation of LRCS Performance

We investigated 147 existing LRCS by analysing the response curves used to characterise the sensitivity of the methods when searching for one left image among $K = 21$ right images, those on the 26 short sequences. The 147 relevant LRCS usable in surgery consist of 143 detector-descriptor combinations using 7 pure local detectors (AGAST, FAST, GFTT, HL, MSD, MSER, STAR), 9 pure local descriptors (BEBLID, BOOST DESC, BRIEF, DAISY, FREAK, LATCH, LUCID, TFEAT, VGG) and 8 local detector-descriptors (AKAZE, BRISK, DISK, KAZE, ORB, SIFT, SuperPoint, SURF). In addition, they include 4 LRCS composed of a global descriptor (GIST), the structural similarity index (SSIM), the video bitrate profile analysis and the deep descriptor (DD) from AlbuNet neural network feature maps initially trained for semantic segmentation of kidney parenchyma. In detail, this model uses the UNet architecture with a Resnet34 encoder pre-trained on ImageNet. For training we created a dataset of 15683 images extracted from 18 RAPN procedures which we manually annotated. We trained with SGD optimiser for 30 epochs with a batch size of 8, a learning rate of 0.001 and a cross-entropy loss function. The ideal response graph would be a peaked bell curve at a 0 offset. Figure 4 and table 1 show that in rapid camera motion sequences (figure 4a, c, g, i), local feature extractors with diffuse spatial support (SIFT, ORB, SURF, SuperPoint), global descriptors (GIST), structural similarity index (SSIM), trained neural network feature maps and the proposed method (SPEC DAISY) reveal sufficient sensitivity to re-synchronise a left-right image pair. However, global methods become unusable in quasi-static scenarios (figure 4b, d, h, j), and methods using diffuse local descriptors exhibit error rates exceeding two frames. The proposed method stands out for its ability to re-synchronise to the nearest frame through its high sensitivity to micro-movements. This sets it apart from diffuse local descriptors, which are inherently insensitive to small movements due to their invariance.

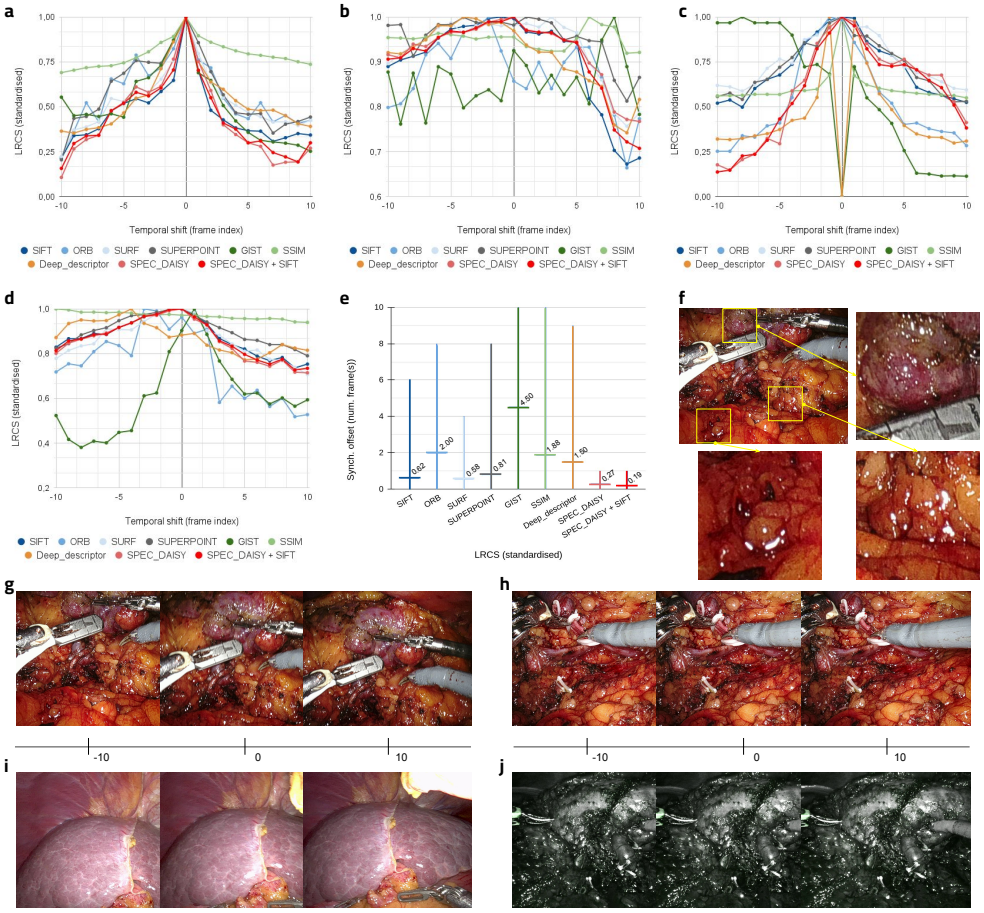


Fig. 4 (a) Response curves for a sequence with rapid camera movement in RAPN. (b) Response curves for quasi-static sequence in RAPN. (c) Response curves for a sequence with rapid camera movement in RAHR. (d) Response curves for quasi-static sequence in RAPN, captured in ICG view mode. (e) Synchronisation offset for the 26 short sequences against the LRCS for local diffuse descriptors (SIFT, ORB, SURF, SuperPoint), global descriptor (GIST), structural similarity index (SSIM), trained neural network feature maps and the proposed methods (SPEC DAISY, SPEC DAISY + SIFT). (f) Sample frame illustrating the size and distribution of specular highlight spots in RAPN. (g) Sample frames of a rapid camera motion sequence in RAPN. (h) Sample frames of a quasi-static sequence in RAPN. (i) Sample frames of a rapid camera motion sequence in RAHR. (j) Sample frames of a quasi-static sequence in RAPN, captured in ICG view mode.

4.3.2 Evaluation of Highlight Detection Impact on Synchronisation

The evaluation of the specular highlight detection module in the SPEC_DAISY framework is crucial for assessing its role in the synchronisation algorithm. To achieve this, a carefully curated dataset of 30 images was created, with an even distribution across robotic-assisted partial nephrectomy, robotic-assisted hepatic resection, and colonoscopy datasets. This compilation is a critical benchmark for scrutinising

Table 1 Synchronisation offset for the 26 short sequences against LRCS for local diffuse descriptors (SIFT, SuperPoint [SP]), trained neural network feature maps [DD] and the proposed methods (SPEC DAISY, SPEC DAISY + SIFT).

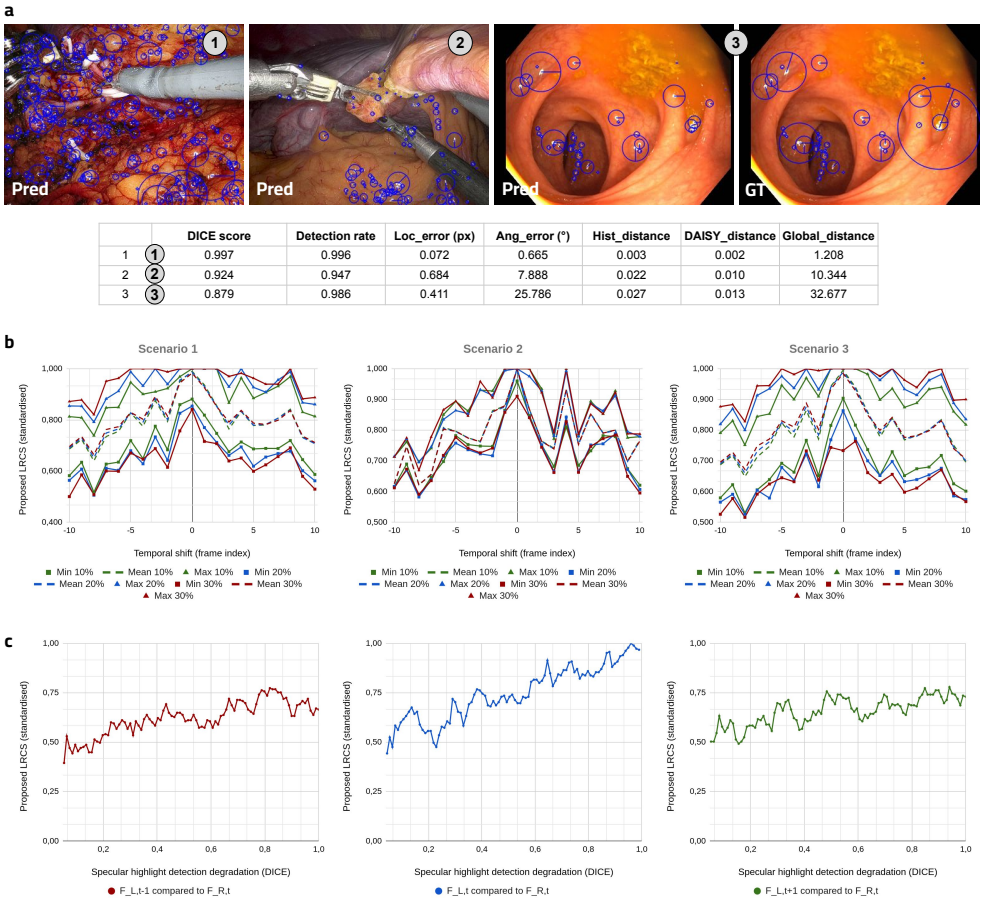
		Synchronisation offset for				
Surgery	Cam. mvt.	SIFT	SP	DD	SPEC_DAISY	SPEC_DAISY + SIFT
RAPN	Slow[1]	0	2	1	0	0
RAPN	None[1]	0	0	1	0	0
RAPN	High[1]	0	1	1	0	0
RAPN	None[1]	0	0	0	0	0
RAPN	None[1]	1	0	0	0	0
RAPN	None[1]	0	1	1	0	0
RAPN	None[1]	0	0	3	0	0
RAPN	High[1]	0	0	0	0	0
RAPN	None[1]	0	0	0	0	0
RAPN	None[1]	0	1	0	1	0
RAPN	None[1]	0	0	0	0	0
RAPN	None[1]	1	1	4	0	0
RAPN	High[1]	0	0	0	0	0
RAPN	High[1]	0	0	0	0	0
RAPN	None[1]	4	3	2	1	1
RAPN	None[1]	0	1	0	0	0
RAPN	None[1]	0	0	7	0	0
RAPN	None[1]	0	1	4	0	0
RAHR	None[1]	6	8	9	1	1
RAHR	None[1]	1	0	0	1	1
RAHR	None[1]	1	1	1	1	1
RAHR	High[1]	0	0	1	0	0
RAPN	None[2]	0	0	1	1	0
RAPN	High[2]	0	0	1	0	0
RAPN	None[2]	0	0	1	0	0
RAPN	None[2]	2	1	1	1	1
	MEAN	0.62	0.81	1.50	0.27	0.19
	STD	1.42	1.65	2.25	0.45	0.40

[1]RGB Camera view

[1]ICG Camera view

the accuracy of specular highlight detection and its impact on the formulation of highlight spot descriptors. For this analytical assessment, we systematically compared the descriptor components extracted from the ground truth images in this dataset with those derived from segmentation masks predicted by the neural network. A comprehensive set of metrics was used, including the detection rate of specular spots, the DICE score, centroid localisation error, angular deviation in the first moment of specular spots, Bhattacharyya distances for histogram congruence, DAISY descriptor distances, and global specular descriptor distances. The results of the study show a $79.85 \pm 23.60\%$ detection rate, with a DICE score of $81.34 \pm 18.11\%$, a centroid localisation error of $1.75 \pm 1.60px$, an angular deviation of $18.45 \pm 12.98°$, a Bhattacharyya distance of 0.06 ± 0.05, a DAISY descriptor distance of 0.02 ± 0.02, and a global specular descriptor distance of 31.36 ± 34.35. These values were computed as the mean and standard deviation across the dataset. Preliminary observations, as shown in figure 5a,

revealed a tendency towards under-segmentation. This can lead to the omission of certain specular spots during the segmentation process.



		DICE score	Detection rate	Loc_error (px)	Ang_error (°)	Hist_distance	DAISY_distance	Global_distance
1	①	0.997	0.996	0.072	0.665	0.003	0.002	1.208
2	②	0.924	0.947	0.684	7.888	0.022	0.010	10.344
3	③	0.879	0.986	0.411	25.786	0.027	0.013	32.677

Fig. 5 (a) Comprehensive evaluation of specular highlight detection in surgical images: a blend of qualitative and quantitative analysis. (b) Graphical depiction of response curves across simulated scenarios 1, 2, and 3. (c) Analysis of synchronisation error in relation to specular highlight detection degradation under sensitivity assessment conditions.

To elucidate the effects of certain phenomena on the synchronisation process, simulations were conducted. These aimed at evaluating the influence of sub-optimal detection and segmentation of specular spots on the synchronisation response curves. Three distinct scenarios were implemented using a short ground truth sequence, comprising 21 images. This raw sequence, numbered 8, is previously referenced in figures 4a. The scenarios are as follows:

- Scenario 1 involves the imprecise detection of a randomly selected 10%, 20%, and 30% subset of specular spots.
- Scenario 2 focuses on the inaccurate segmentation of a randomly selected 10%, 20%, and 30% subset of specular spots.
- Scenario 3 combines both sub-optimal detection and segmentation of a randomly chosen 10%, 20%, and 30% subset of specular spots.

The response curves, depicted in figure 5b, were obtained by executing each of the 9 experiments 1000 times. Furthermore, figure 4c includes the synchronisation error under sensitivity evaluation conditions (the response curves) as a function of specular highlight detection degradation. These results suggest that synchronisation is minimally affected as long as highlight detection specificity and repeatability is maintained. To guarantee the system's capacity to re-synchronise stereo video streams, it is crucial to adhere to a response criterion greater than 77% for detecting subpar specular spots. Our experiments demonstrate that the proposed neural network architecture exceeds this threshold, validating its effectiveness under challenging conditions.

4.3.3 Evaluation of LRCS Impact on Synchronisation

The synchronisation method's performance is examined in detail on 18 de-synchronised sequences, hence 3600 frames generated by the de-synchronisation simulator based on the extended sequences of the raw dataset and with parameters suitable for the 60 Hz stream. Without temporal smoothing, we achieve an average offset of 0.055 frames and a maximum offset of 1 frame for the proposed specular LRCS, with $94.1 \pm 3.6\%$ of frames successfully synchronised. The top-performing existing LRCS achieves an average and maximum offset of 0.3 and 3 frames, with $81.2 \pm 6.4\%$ successfully synchronised frames. The use of the proposed diffuse-specular LRCS and EWMA as smoothing function increases performance to $98.6 \pm 0.6\%$ of successfully synchronised frames. The time taken to process each new pair of incoming images, from receiving the images to outputting the synchronised pair, including all stages of image detection, description, matching, and selection, is lower than 14 ms using multithreading on an Ubuntu PC with Intel Core™ i9-10900K and an Nvidia RTX 2080Ti GPU card, well below the 60 Hz stream frame rate.

A statistical study aims to evaluate the effectiveness of the proposed specular LRCS or the diffuse-specular LRCS (Method A) and an existing top-performing LRCS (Method B) is performed. The comparison focuses on two main metrics: the offset in synchronisation (both average and maximum) and the percentage of successfully synchronised frames. The Null Hypothesis (H0) is "There is no significant difference between the proposed method and the baseline method in terms of average offset, maximum offset, and synchronisation success rate". The Alternative Hypothesis (H1) is "There is a significant difference between the proposed method and the baseline method in these metrics". For analysis, a common threshold for significance is set at 0.05. If $p < 0.05$, the null hypothesis is rejected, suggesting a significant difference between the methods. If $p \geq 0.05$, the null hypothesis is not rejected, suggesting no significant difference. In this study, paired statistical methods were used to evaluate the same video streams with both Method A and Method B, facilitating direct performance

comparison. A Paired T-test analysed the synchronisation offsets, while a Z-test for proportions assessed the success rates, with respective p-values 5.47×10^{-9} and 9.10×10^{-14} indicating the significance of the results. These techniques provided detailed insights into the performance differences, confirming Method A's statistical superiority over Method B.

4.4 Application

The proposed synchronisation system was developed for the surgical context, optimising data synchrony for CAS systems, whether for camera calibration, intraoperative 3D reconstruction or stereo organ tracking. A clear example of the benefits of accurate synchronisation for depth estimation can be seen in figure 6a. The proposed LRCS ensures synchronisation between the left and right channels, resulting in accurate and coherent disparity maps. A delay of one frame lasting approximately 17 ms results in several inaccuracies (figure 6b) in severe disparity computation, using SGBM as an example of depth estimation.

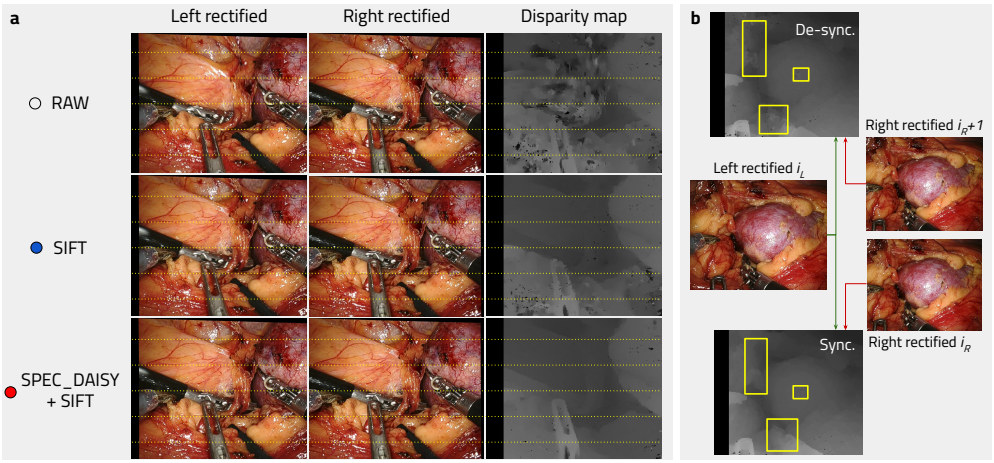


Fig. 6 Qualitative assessment of SGBM depth estimation on a) raw desynchronised real stereo images showing very poor disparity estimation. Synchronisation methods based on diffuse reflection (SIFT) improve the quality of the estimate. The addition of specular reflection (SPEC DAISY + SIFT) gives the best qualitative results with very few errors. b) In a semi-synthetic case, a single image synchronisation error generates multiple disparity estimation errors (yellow rectangles).

5 Conclusion

We have proposed a pipeline to synchronise surgical stereoscopic video streams. Our pipeline takes full advantage of laparoscopic images by exploiting both diffuse and specular reflections with the new Diffuse-Specular LRCS. The use of specular reflection brings a tremendous boost to the real-time surgical stereo synchronisation problem

due to its high discriminative properties. Currently, the method is primarily limited by the amount of specular blobs detected on moist tissue surfaces.

In future work, we will focus on integrating the proposed method to a CAS system for partial nephrectomy and testing in the operating theatre, to demonstrate the effectiveness of the real-time system in recovering reliable stereo-synchronised video streams.

Acknowledgments

We express our gratitude to UroCCR: French Kidney Cancer Research Network and the Clermont-Ferrand University Hospital in France for supplying the anonymised surgical videos used in this study. This work was supported by the French Government through the France 2030 program (ANR-21-RHU5-0015).

Declarations

Conflict of Interest: The authors declare that they have no conflict of interest. **Ethical approval:** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors. **Informed consent:** Informed consent was obtained from the patients included in the study.

References

[1] Vittori, G.: Open versus robotic-assisted partial nephrectomy: a multicenter comparison study of perioperative results and complications. World Journal of Urology **32**, 287–293 (2014)

[2] Khaddad, A., Bernhard, J.C., Margue, G., Michiels, C., Ricard, S., Chandelon, K., Bladou, F., Bourdel, N., Bartoli, A.: A survey of augmented reality methods to guide minimally invasive partial nephrectomy. World Journal of Urology **41**, 335–343 (2022)

[3] Roberts, S.I., Desai, A., Checcucci, E., Puliatti, S., Taratkin, M.S., Kowalewski, K.F., Rivas, J.G., Rivero, I., Veneziano, D., Autorino, R., Porpiglia, F., Gill, I.S., Cacciamani, G.E.: "augmented reality" applications in urology: a systematic review. Minerva urology and nephrology (2022)

[4] Hou, L., Kagami, S., Hashimoto, K.: Frame synchronization of high-speed vision sensors with respect to temporally encoded illumination in highly dynamic environments. Sensors (Basel, Switzerland) **13**, 4102–4121 (2013)

[5] Froner, D.S., Pio, J.L.S., Simões, W.C.S.S.: Spatiotemporal alignment of sequential images: Utilization of gps data as reference to alignment of multiple video cameras. 2016 XLII Latin American Computing Conference (CLEI), 1–7 (2016)

[6] Ushizaki, M., Okatani, T., Deguchi, K.: Video synchronization based on co-occurrence of appearance changes in video sequences. 18th International Conference on Pattern Recognition (ICPR'06) **3**, 71–74 (2006)

[7] Melloni, A., Lameri, S., Bestagini, P., Tagliasacchi, M., Tubaro, S.: Near-duplicate detection and alignment for multi-view videos. 2015 IEEE International Conference on Image Processing (ICIP), 2444–2448 (2015)

[8] Dexter, E., Pérez, P., Laptev, I., Junejo, I.N.: View-independent video synchronization from temporal self-similarities. In: International Conference on Computer Vision Theory and Applications (2009)

[9] Dai, C., Zheng, Y., Li, X.: Subframe video synchronization via 3d phase correlation. 2006 International Conference on Image Processing, 501–504 (2006)

[10] Pinheiro, G., Cirne, M.V.M., Bestagini, P., Tubaro, S., Rocha, A.: Detection and synchronization of video sequences for event reconstruction. 2019 IEEE International Conference on Image Processing (ICIP), 4060–4064 (2019)

[11] Polok, L., Klicnar, L., Beran, V., Smrž, P., Zemčík, P.: Quality assurance in large collections of video sequences. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 3580–3584 (2015). IEEE

[12] Cao, X., Xiao, J., Foroosh, H.: Camera motion quantification and alignment. 18th International Conference on Pattern Recognition (ICPR'06) **2**, 13–16 (2006)

[13] Diego, F., Serrat, J., Peña, A.M.L.: Joint spatio-temporal alignment of sequences. IEEE Transactions on Multimedia **15**, 1377–1387 (2013)

[14] Caspi, Y., Irani, M.: A step towards sequence-to-sequence alignment. Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662) **2**, 682–6892 (2000)

[15] Caspi, Y., Irani, M.: Spatio-temporal alignment of sequences. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 1409–1424 (2002)

[16] Sinha, S.N., Pollefeys, M.: Camera network calibration and synchronization from silhouettes in archived video. International Journal of Computer Vision **87**, 266–283 (2010)

[17] Albl, C., Kukelova, Z., Fitzgibbon, A.W., Heller, J., Smíd, M., Pajdla, T.: On the two-view geometry of unsynchronized cameras. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5593–5602 (2017)

[18] Lee, S.-Y., Sim, J.-Y., Kim, C.-S., Lee, S.U.: Correspondence matching of multi-view video sequences using mutual information based similarity measure. IEEE Transactions on Multimedia **15**, 1719–1731 (2013)

[19] Schweiger, F., Schroth, G., Eichhorn, M., Al-Nuaimi, A., Cizmeci, B., Fahrmair, M., Steinbach, E.G.: Fully automatic and frame-accurate video synchronization using bitrate sequences. IEEE Transactions on Multimedia **15**, 1–14 (2013)

[20] Tola, E., Lepetit, V., Fua, P.: Daisy: An efficient dense descriptor applied to wide-baseline stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**, 815–830 (2010)

[21] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**, 91–110 (2004)