Yanagawa *et al. IPSJ Transactions on Computer Vision and Applications* (2017) 9:3
DOI 10.1186/s41074-017-0015-6

IPSJ Transactions on Computer Vision and Applications

EXPRESS PAPER Open Access

Abnormality tracking during video capsule endoscopy using an affine triangular constraint based on surrounding features

Yukiko Yanagawa[1*], Tomio Echigo[2†], Hai Vu[3], Hirotoshi Okazaki[4], Yasuhiro Fujiwara[4], Tetsuo Arakawa[4] and Yasushi Yagi[5†]

Abstract

The precise tracking of an abnormality in the gastrointestinal tract is useful for medical training purposes. However, the gastrointestinal wall deforms continuously in an unpredictable manner, while abnormalities lack distinctive features, making them difficult to track over continuous frames. To address this problem, we propose a tracking method for capsule endoscopy using the surrounding features of abnormalities. By applying triangular constraints using an affine transformation, we are able to track abnormalities that do not have distinctive features over consecutive image frames. We demonstrate the efficacy of our approach using eight common types of gastrointestinal abnormalities.

Keywords: Object tracking, Deformable object, Wireless capsule endoscopy images, Education system, Gastrointestinal

1 Introduction

Medical endoscopes are widely used in procedures for inspecting the inner cavities of the human body. In gastrointestinal (GI) surgery, endoscopy is often favored over laparotomy because it is less stressful for the patient and is less labor-intensive for the surgeon. GI mapping that identifies the position of an abnormal region, once performed manually, is now being carried out using medical imaging. In this process, a three-dimensional map is constructed from a sequence of two-dimensional endoscopy images captured during GI treatment. Feature tracking and image analysis in GI mapping have been the subject of a large body of research [1–3].

Wireless capsule endoscopy (WCE), which has been in use since 2002, employs a device that is swallowed and propelled by peristalsis through the GI tract. The capsule captures images at a low frame rate during its 7–8-h passage through the GI tract. A WCE diagnosis requires doctors to have skills different from those required for conventional endoscopy. Abnormalities in WCE images do not have clear edges or present with appreciable contrast to the background tissue, as seen in the examples in Fig. 1. There has thus been much research centered on assisting WCE diagnosis in recent years [4–8]. Furthermore, a training system is needed to assist medical doctors to identify abnormal regions in WCE images.

When a medical doctor unskilled in WCE identifies an abnormality on an image in a training system, ground truth data, which include the period between the abnormality's appearance and disappearance, as well as the abnormality's type and location, are needed to decide whether the identification is correct. Currently, medical doctors manually annotate each frame. However, as an abnormality can appear in any number of frames, ranging from only a few to several hundred, this is a time-consuming task. There is clearly a need to be able to automatically track the abnormality.

The present study focuses on abnormality tracking in a sequence of WCE images. Clearly, as the shape of the GI tract is deformed by peristaltic motion, the shape of the abnormality changes in these images (Figs. 2 and 3). The boundaries of abnormal regions, such as bleeding sites and tumors (Fig. 1), are difficult to distinguish using standard image feature detectors. Moreover, the position of such extractable features is unstable even if the change

*Correspondence: yanagawa@ari.ncl.omron.co.jp
†Equal Contributors
[1]OMRON Corporation, 9-1, Kizugawadai, Kizugawa-shi, Kyoto, Japan
Full list of author information is available at the end of the article

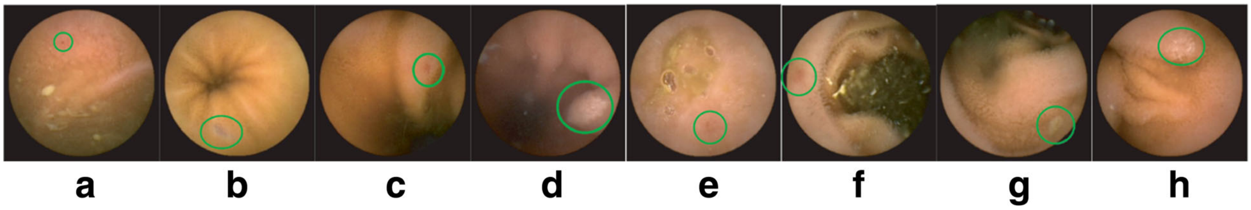


Fig. 1 Samples of GI abnormalities within the small intestine. **a** *Red spot*. **b** Phlebectasia. **c** Angiodysplasia. **d** Lymphangiectasia. **e** Erosion. **f** Erythematous. **g** Ulcer. **h** White-tipped villi. *Green circles* in each image identify the abnormal regions

between consecutive images is small. The main cause of failure in algorithmic GI abnormality tracking is the lack of distinguishable features in an abnormality target region.

Nontrivial image changes, related to the low frame rate of the video capture, also occur during WCE. Motion prediction approaches, such as the use of the Kalman filter [9, 10] and particle filter [11], cannot be used because image changes caused by peristaltic motion appear to be random. Recently, methods of tracking based on convolutional neural networks (CNNs) have been proposed [12, 13]; CNN methods have produced excellent results [14] but require a lot of data for training.

In our proposed method, we use features of the abnormality, as well as surroundings features, or supporters. Generally, supporters and the target are inherently difficult to track using their ambiguous image features alone. We thus employ three mutual supporters with a triangular constraint, which preserves a triangular shape but allows weak deformation between consecutive frames. We found that the position of the target could be reliably determined using the transient position over successive frames, on the basis of the affine transformation of several supporters.

Kanazawa and Uemura [15] proposed a method for matching weak features between two affine invariant images, using a triplet vector descriptor; however, their method does not support images that have local changes. To address this limitation, we placed constraints on each matching supporter but not on the image as a whole, allowing the matching of deformable objects. Grabner et al. [16] developed a method of tracking invisible targets using the surrounding features, such as the large eigenvalues of a Hessian matrix. The method typically requires that the target has strong image features. Here, we propose a tracking procedure for abnormalities using a triangular constraint depending on the angle method based on the surrounding features [17]. This procedure is designed for tracking deformable and weak-featured objects. However, it is not enough to simply match pairs using this procedure, because angular constraints result in a high flexibility of matching pairs. We thus focused on the wall of the GI tract, using local affine constraints to track an abnormality, on the basis of its surrounding features.

2 Method

2.1 Overview

We assume that a medical doctor or clinical technician identifies the initial location of an abnormal region within an image through manual analysis. Starting from a given frame, our method can track the abnormal region, forwards and backwards, to determine its appearance period. Our method involves three stages, outlined below.

First, feature points in the target (abnormal region) and in the surrounding features (supporters) between successive images are matched. Usually, supporters cannot be accurately tracked individually, because an image sequence from a capsule video does not have clear features. For this reason, a triangular constraint is used, which follows the relative positions across consecutive frames.

Second, the target position is roughly estimated using a voting process based on an affine matrix that is calculated from triplet supporters. The relative position of the

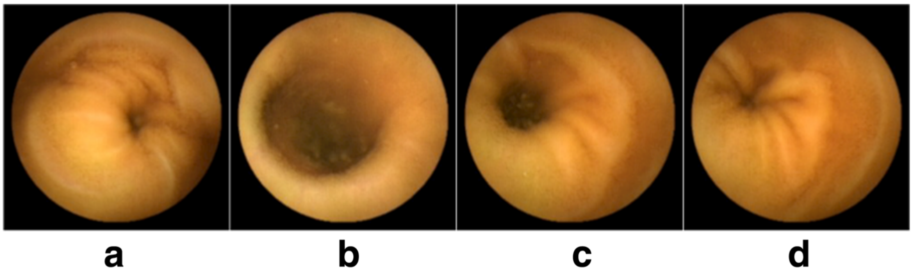


Fig. 2 a–**d** Example of four continuous frames in a capsule endoscopy image sequence

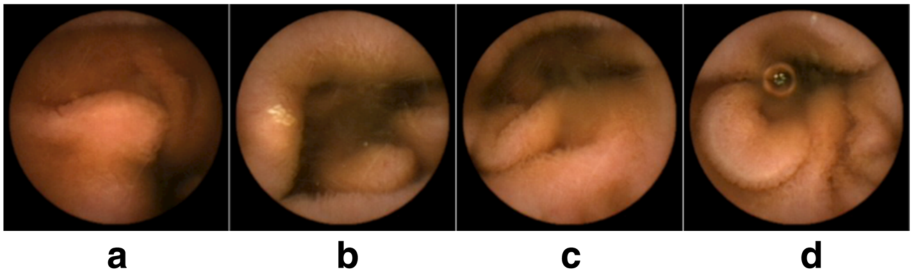


Fig. 3 a–**d** Example of four continuous frames in a capsule endoscopy image sequence

target is adjusted using a mapping between triangles of the current and subsequent frames.

Finally, a precise position is estimated using features inherent to the abnormality. This involves correcting for errors introduced during the coarse estimation in the previous step, usually caused by deformation-based movement of the scene. For this process, we use color information from the target, rather than traditional anchor points, as this is more reliable when there are several independent moving regions. We derive the theory behind these three stages below.

2.2 Matching supporters

In the first stage, supporter pair sets $\mathbf{P}^t$ are created between consecutive frames t and $t-1$. In each frame, the supporters are extracted using the Kanade-Lucas-Tomasi (KLT) method [18]. In the proposed method, the KLT method is used only for point detection and not for point matching. Our proposed method does not depend on this point detection method if enough points are detected in continuous frames. Two other point detection methods were also evaluated: the scale-invariant feature transform (SIFT)[19] and speeded-up robust features (SURF)[20] methods. The target region in the small intestine, used as a test case here, does not have a defining texture; hence, SURF and SIFT, both being local to region-based feature descriptors, did not fare well. The KLT method, however, is normally defined by 3×3 pixels and is sensitive to changes in color intensity. The KLT method was found to be more stable and accurate for matching points of the bowel wall.

A supporter pair set $\mathbf{P}^t$ is given by

$$\mathbf{P}^t = \left\{\mathbf{p}^t_{(i,l)} \mid i \in N^t_p, l \in N^{t-1}_p\right\}, \tag{1}$$

where N^t_p and N^{t-1}_p are the numbers of supporters in frames t and $t-1$. The support pairs

$$\mathbf{p}^t_{(i,l)} = \left(\mathbf{F}^t_i, \mathbf{F}^{t-1}_l\right), \tag{2}$$

have the elements

$$\mathbf{F}^t_i = \left(\mathbf{x}^t_i, \boldsymbol{f}^t_i\right), \tag{3}$$

$$\boldsymbol{f}^t_i = \left\{\mathbf{h}^t_{i1}, \mathbf{h}^t_{i2}, \cdots, \mathbf{h}^t_{iH}\right\}, \tag{4}$$

where $\mathbf{P}^t$ has $N^t_p \times N^{t-1}_p$ supporter pairs that are a combination of all detected points for frames t and $t-1$, the supporter $\mathbf{F}^t_i$ is the ith point in frame t, $\mathbf{x}^t_i$ is the coordinate of point i in frame t, and $\boldsymbol{f}^t_i$ is the image feature, denoted by components of the GI color histogram (Eq. 4) [21] around KLT feature points. We found that the GI color space was an efficient means of describing color features in a sequence of WCE images. The GI color histogram was proposed to better distinguish suspicious regions from normal regions without overly enhancing the image. The GI color histogram is constructed employing principal component analysis from a large dataset of capsule endoscopy sequences that covers a variety of patient data. The GI color histogram uses the third component. H is the number of partitions in the histogram. $\mathbf{h}^t_{ia}$ denotes the value of the ath partition in the histogram. $\boldsymbol{f}^t_i$ is the normalized histogram. Each frame is divided into four areas, and the supporters are detected in each area. The same number of supporters is detected in each area and this number is set in advance. A supporter pair $\mathbf{p}^t_{(i,l)}$ in the supporter pair set $\mathbf{P}^t$ is matched according to its matching score $\mathbf{Bc_f}(\mathbf{p}^t_{(i,l)})$ and its weighting $\mathbf{Mw}(\mathbf{p}^t_{(i,l)})$ from the triangular constraint. The score $\mathbf{Bc_f}(\mathbf{p}^t_{(i,l)})$ denotes the Bhattacharyya distance between features $\boldsymbol{f}^t_i$ and $\boldsymbol{f}^{t-1}_l$ and is given by

$$\mathbf{Bc_f}\left(\mathbf{p}^t_{(i,l)}\right) = \sum_{a=1}^{H} \sqrt{h^t_{i\,a} h^{t-1}_{l\,a}}. \tag{5}$$

Additionally, the matching score flag is defined by

$$\mathbf{J_f}\left(\mathbf{p}^t_{(i,l)}\right) = \begin{cases} 1, \ if \ \mathbf{Bc_f}\left(\mathbf{p}^t_{(i,l)}\right) \geq \mathbf{Th_B}, \\ 0, \ else, \end{cases} \tag{6}$$

The WCE images capture localized movements in different regions between successive frames. We use a triangular constraint weight $\mathbf{Mw}(\mathbf{p}^t_{(i,l)})$ to match supporters, which maintains a triangular shape between successive frames. In this way, we can assume that successive images represent a rigid state.

We now need to define the scoring algorithm for supporter pair $\mathbf{p}^t_{(i,l)}$, stored in the supporter pair sets $\mathbf{P}^t$. The triangular constraint weight of a supporter pair $\mathbf{p}^t_{(i,l)}$ is

$$\mathbf{Mw}\left(\mathbf{p}^t_{(i,l)}\right) = \mathbf{J_f}\left(\mathbf{p}^t_{(i,l)}\right) \frac{\sum_j \sum_m \sum_k \sum_n \mathbf{J_f}\left(\mathbf{p}^t_{(j,m)}\right) \mathbf{J_f}\left(\mathbf{p}^t_{(k,n)}\right) \mathbf{DM}\left(\mathbf{p}^t_{(i,l)}, \mathbf{p}^t_{(j,m)}, \mathbf{p}^t_{(k,n)}\right)}{\sum_j \sum_m \sum_k \sum_n \mathbf{J_f}\left(\mathbf{p}^t_{(j,m)}\right) \mathbf{J_f}\left(\mathbf{p}^t_{(k,n)}\right)}, \tag{7}$$

$$j = [1, 2, \cdots, N^t_p - 1,\ j \neq i], \tag{8}$$

$$k = [j+1, j+2, \cdots, N^t_p,\ k \neq i], \tag{9}$$

$$m = [1, 2, \cdots, N^{t-1}_p - 1,\ m \neq l], \tag{10}$$

$$n = [m+1, m+2, \cdots, N^{t-1}_p,\ n \neq l], \tag{11}$$

where

$$\mathbf{DM}\left(\mathbf{p}^t_{(i,l)}, \mathbf{p}^t_{(j,m)}, \mathbf{p}^t_{(k,n)}\right) = \begin{cases} 0, \\ \text{if either the triplet supporters } \left(\mathbf{F}^t_i, \mathbf{F}^t_j, \mathbf{F}^t_j\right) \text{ or } \left(\mathbf{F}^{t-1}_l, \mathbf{F}^{t-1}_m, \mathbf{F}^{t-1}_n\right) \text{ is collinear.} \\ 1, \\ \text{if all parameters of the affine matrix } \mathbf{A}^t_{(i,l),(j,m),(k,n)} \text{are within acceptable value.} \\ 0,\ \textit{else} \end{cases}$$

$$\mathbf{p}^t_{(j,m)} = \left(\mathbf{F}^t_j, \mathbf{F}^{t-1}_m\right), \tag{12}$$

$$\mathbf{p}^t_{(k,n)} = \left(\mathbf{F}^t_k, \mathbf{F}^{t-1}_n\right). \tag{13}$$

The supporter pairs $\mathbf{p}^t_{(i,l)}, \mathbf{p}^t_{(j,m)}, \mathbf{p}^t_{(k,n)}$ are used to compute an affine matrix:

$$\mathbf{A}^t_{(i,l),(j,m),(k,n)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} = \mathbf{A_{Scale}} * \mathbf{A_{Rotation}} * \mathbf{A_{Shear}} * \mathbf{A_{Translation}}, \tag{14}$$

where

$$\mathbf{A_{Scale}} = \begin{bmatrix} S_x & 0 & 0 \\ 0 & S_y & 0 \end{bmatrix}, \tag{15}$$

$$\mathbf{A_{Rotation}} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \end{bmatrix}, \tag{16}$$

$$\mathbf{A_{Shear}} = \begin{bmatrix} 1 & \tan\alpha_y & 0 \\ \tan\alpha_x & 1 & 0 \end{bmatrix}, \tag{17}$$

$$\mathbf{A_{Translation}} = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \end{bmatrix}. \tag{18}$$

The elements a_{13} and a_{23} are defined by the translation parameters (t_x, t_y), and the remaining elements a_{11}, a_{12}, a_{21}, and a_{22} are the result of rotation through angle θ, scaling by (S_x, S_y), and shearing at angles (α_x, α_y). The affine matrix $\mathbf{A}^t_{(i,l),(j,m),(k,n)}$ is defined by a rotational angle θ, scale changes (S_x, S_y), shearing angles (α_x, α_y), and translational parameters (t_x, t_y), as shown in Eqs. 14–18.

The acceptable limits for these parameter values in our application are

- $-20° <$ rotational angle $\theta < 20°$
- $-10° <$ shearing angle $< 10°$
- $0.25 <$ scale change < 3.0
- Maximum translational range, 1/2 image size

These parameters were empirically determined in our study since a_{11}, a_{12}, a_{21}, and a_{22} cannot be decomposed into θ, (S_x, S_y), (α_x and α_y). Whether these values of a_{11}, a_{12}, a_{21}, and a_{22} are within possible ranges is calculated in advance. The angular constraint that was used in our previous work [17] is only a limit of the shearing angle and scale change.

Our method for selecting which supporter pairs are stored in selected supporter pair sets $\mathbf{G}^t$ is as follows.

This involves three steps.

Step 1: Calculate the score $\mathbf{Zc}^t_{(i,l)}$ for all supporter pairs $\mathbf{p}^t_{(i,l)}$ in set $\mathbf{P}^t$,

$$\mathbf{Zc}^t_{(i.l)} = \mathbf{Mw}\left(\mathbf{p}^t_{(i,l)}\right) \mathbf{Bc_f}\left(\mathbf{p}^t_{(i,l)}\right) \tag{19}$$

Step 2: The supporter pair $\mathbf{p}^t_{(i,l)}$ having the highest score $\mathbf{Zc}^t_{(i,l)}$ in set $\mathbf{P}^t$ is selected and stored in selected supporter pair set $\mathbf{G}^t$.

Step 3: Supporter pairs that include points i and l are rejected from the set $\mathbf{P}^t$.

Return to step 2.

This process is repeated until all supporter pairs are rejected. The selected supporter pair set $\mathbf{G}^t$ is a subset of supporter pair set $\mathbf{P}^t$:

$$\mathbf{G}^t \subset \mathbf{P}^t \tag{20}$$

$$\mathbf{G}^t = \left\{\mathbf{p}^t_s \mid s \in N^t_g\right\}. \tag{21}$$

N^t_g is then the number of selected supporter pairs.

Figure 4 shows the point matching results obtained without a constraint and using an angular constraint [17] and our proposed method. Blue points represent detected points while red lines match saturation scores (where a high saturation corresponds to a high score).

2.3 Coarse approximation of an abnormal region

In this stage, we produce a rough estimate of the abnormality's position. After creating supporter pairs between successive frames, a voting map is created. Abnormalities are located on the voting map using an affine matrix that is calculated from each matching supporter triplet. Conventionally, a one-to-one relationship is used in the voting process for the target position, depending on the parallel shift of the camera (Fig. 4a). This process cannot,

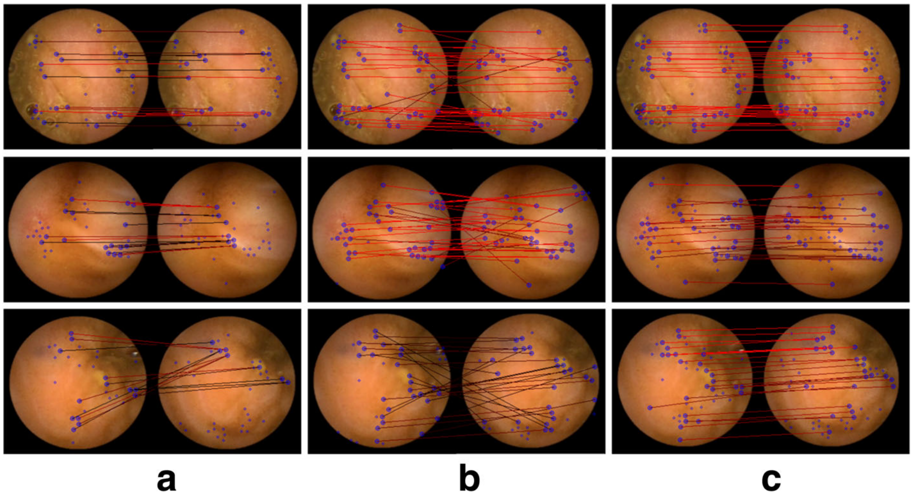


Fig. 4 Point matching results using **a** no constraint (the conventional method), **b** an angular constraint [17], and **c** an affine constraint (the proposed method)

however, cope with any rotational movement of the camera (Fig. 4b, c). To overcome this limitation, our method uses relationships between the triplets of supporters and the abnormality's position, when voting for the target position. Figure 5 shows the voting results for the conventional and proposed methods. The red rectangle gives the actual abnormal region, while the green rectangle gives the abnormal region estimated using the conventional method and the blue rectangles give the abnormal regions estimated using our proposed method. It is clear that our approach performs better than the conventional method. Our approximation of position, based on a voting map, is outlined in detail below.

The score of each (x, y) position on the voting map is given by

$$\mathbf{C}(x,y) = \sum_{u>t}^{N_g^t} \sum_{t>s}^{N_g^t-1} \sum_{s=1}^{N_g^t-2} \mathbf{f}(x,y)\mathbf{Bc_f}\left(\mathbf{p}_s^t\right)\mathbf{Bc_f}\left(\mathbf{p}_t^t\right)\mathbf{Bc_f}\left(\mathbf{p}_u^t\right), \tag{22}$$

where

$$\mathbf{f}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{1}{2}\left(\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}\right)\right) \tag{23}$$

is the weight of the distance to the estimated position of the abnormality, produced using the affine matrix $\mathbf{A}_{s,t,u}^t$. The tracking area is defined as a rectangle with sides parallel to the x axis and y axis. $(\sigma_\mathbf{x}, \sigma_\mathbf{y})$ denotes the lengths of the sides of the rectangle. The voting scores $\mathbf{C}(x, y)$ are calculated by multiplying the three matching scores $\mathbf{Bc_f}(\cdot)$ and the weighting function $\mathbf{f}(x, y)$, which is defined by the distance from the abnormality's position, estimated from the triplet supporter pairs that are stored in the selected supporter pair set $\mathbf{G}^t$.

The estimated position of the abnormality in the current frame, $(\mu_\mathbf{x}, \mu_\mathbf{y})$, is computed using the affine transformation of the previous abnormality's position (t_x, t_y), i.e.,

$$\begin{bmatrix}\mu_x\\ \mu_y\\ 1\end{bmatrix} = \mathbf{A_{s,t,u}}\begin{bmatrix}t_x\\ t_y\\ 1\end{bmatrix}. \tag{24}$$

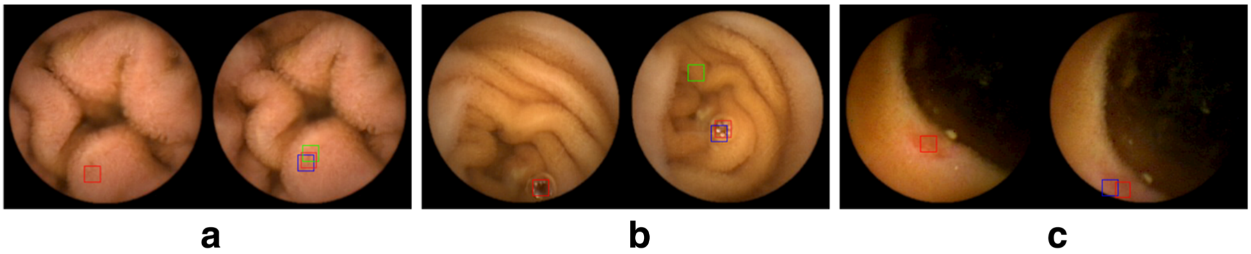


Fig. 5 Movement between successive frames caused by **a** a parallel shift and **b**, **c** rotational movements of the camera. The *red rectangles* identify abnormal regions. The *green rectangles* show the attempt to track these abnormalities by assuming only a parallel shift (the conventional method) whereas the *blue rectangles* show the attempt to track these abnormalities using affine voting (the proposed method)

2.4 Determining a precise position

We now explain how we track an abnormal region. The voting procedure can result in an error in the position of the abnormal region, because of continuous deformation of the scene. Hence, we now use color information of the target, which is more reliable in situations that the target is deformable. In the initial frame, an abnormal region is described using the GI color space, which was developed to distinguish abnormal regions within the intestine [21]. We found the third GI color component to be the most useful for this purpose.

Next, a score is calculated for all positions having a voting score greater than the threshold:

$$\mathbf{V}(x,y) = \mathbf{C}(x,y)\mathbf{B}(Ab_{\text{initial}}, F),$$

$$\mathbf{V}(x,y) = \mathbf{C}(x,y)\mathbf{Bc_a}\left(\boldsymbol{Q_{initial}}, \boldsymbol{K}_{(\boldsymbol{x},\boldsymbol{y})}\right) \tag{25}$$

$$\mathbf{Bc_a}\left(\boldsymbol{Q_{initial}}, \boldsymbol{K}_{(\boldsymbol{x},\boldsymbol{y})}\right) = \sum_{a=1}^{H} \sqrt{q^{t}_{i\,a} k^{t-1}_{l\quad a}}, \tag{26}$$

where (x,y) is the position, $C(x,y)$ is the voting score calculated from the coarse estimation, and $\mathbf{Bc_a}(\boldsymbol{Q_{initial}}, \boldsymbol{K}_{(\boldsymbol{x},\boldsymbol{y})})$ is the Bhattacharyya distance between a GI color normalized histogram of the target in the initial frame ($\boldsymbol{Q_{initial}}$) and that around position (x,y) in the current frame ($\boldsymbol{K}_{(\boldsymbol{x},\boldsymbol{y})}$). The position having the maximum score is designated as the target. If the maximum $\mathbf{Bc_a}(\boldsymbol{Q_{initial}}, \boldsymbol{K}_{(\boldsymbol{x},\boldsymbol{y})})$ is lower than the threshold and the voting score in the image area is less than or equal to the threshold, then the current frame does not include the target.

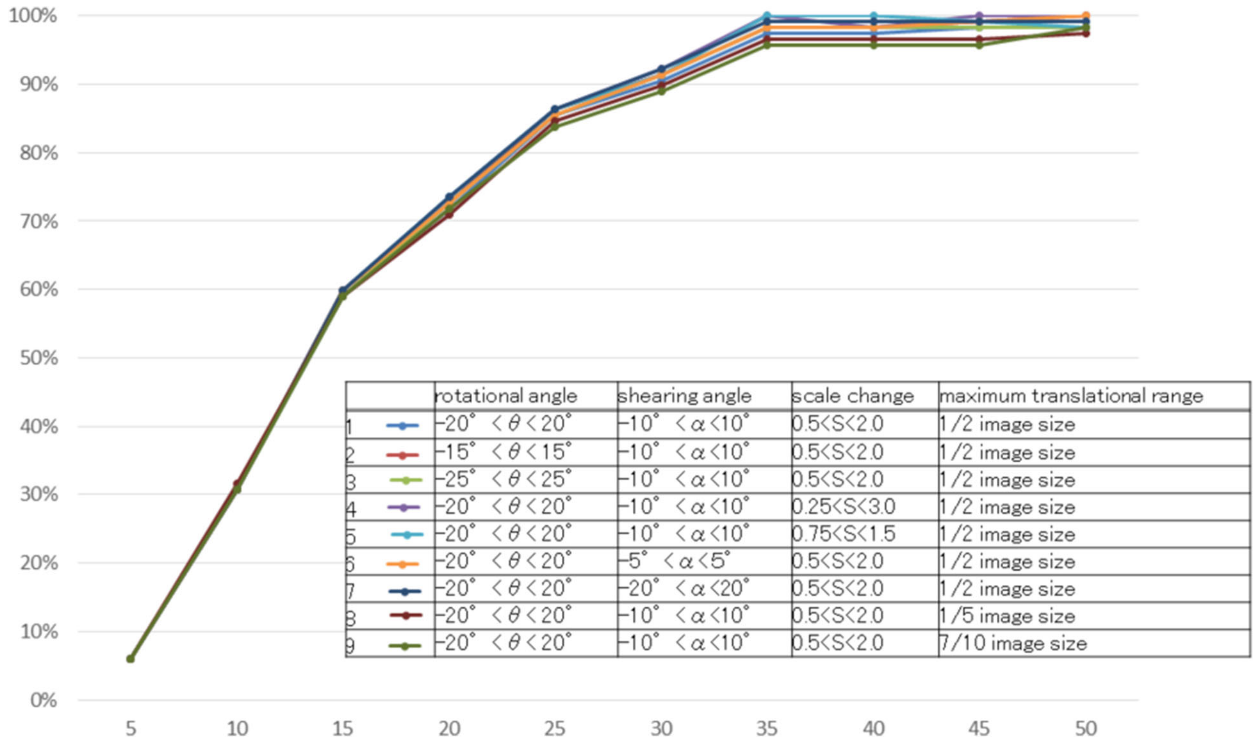
3 Results

We processed 120 sequences, chosen randomly, in which eight major types of abnormalities were present: a red spot, phlebectasia, angiodysplasia, lymphangiectasia, erosion, erythematous, ulcer, and white-tipped villi (see Fig. 1). The image size was 256 × 256 pixels and images were recorded at 2 fps. The ground truth of the abnormal position and size were demarcated manually. If the size of the abnormality was smaller than 30 pixels, then the size was set to 30 pixels. Figure 6 shows the precision and success plots per sequence obtained using the proposed method, i.e., changing each limited parameter of the affine matrix in the stage of matching supporters.

The tracking error in frame t, denoted $\mathbf{E}^t$, is defined as

$$\mathbf{E}^t = \sqrt{(x_{e^t} - x_{a^t})^2 + (y_{e^t} - y_{a^t})^2}, \tag{27}$$

where (x_{e^t}, y_{e^t}) denotes the central position of estimated tracking area in frame t and (x_{a^t}, y_{a^t}) denotes the central position of the ground truth of the abnormal area in frame



		rotational angle	shearing angle	scale change	maximum translational range
1	—●—	−20° < θ < 20°	−10° < α <10°	0.5<S<2.0	1 /2 image size
2	—●—	−15° < θ < 15°	−10° < α <10°	0.5<S<2.0	1 /2 image size
3	—●—	−25° < θ < 25°	−10° < α <10°	0.5<S<2.0	1 /2 image size
4	—●—	−20° < θ < 20°	−10° < α <10°	0.25<S<3.0	1 /2 image size
5	—●—	−20° < θ < 20°	−10° < α <10°	0.75<S<1.5	1 /2 image size
6	—●—	−20° < θ < 20°	−5° < α <5°	0.5<S<2.0	1 /2 image size
7	—●—	−20° < θ < 20°	−20° < α <20°	0.5<S<2.0	1 /2 image size
8	—●—	−20° < θ < 20°	−10° < α <10°	0.5<S<2.0	1 /5 image size
9	—●—	−20° < θ < 20°	−10° < α <10°	0.5<S<2.0	7/10 image size

Fig. 6 Precision and success plots per sequence. Here, the performance of the proposed method is investigated by changing parameters

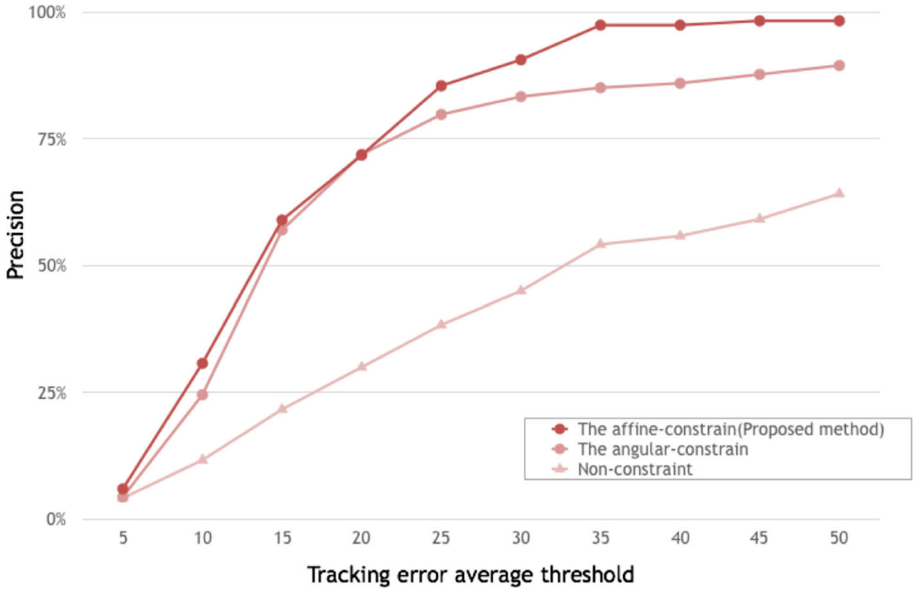


Fig. 7 Precision and success plots per sequence. Here, the performance of the proposed method is compared with that of the conventional method and angular-constraint method

t. It was found that the proposed method is insensitive to parameter changes. We see that parameter set 4 provides better results than other parameter sets. The parameters were defined in the method section.

Figure 7 shows precision and success plots per sequence when using the proposed method and the angular constraint [17] and no constraint (i.e., conventional method).

We see that our proposed method performs better than the other methods and is able to track an abnormality, with an error of less than 30 pixels, for more than 90% of the sequence.

Figure 8 compares the proposed method with both the mean-shifted method [22] and the multi-domain network (MDNet) [12] method that won the Visual Tracking

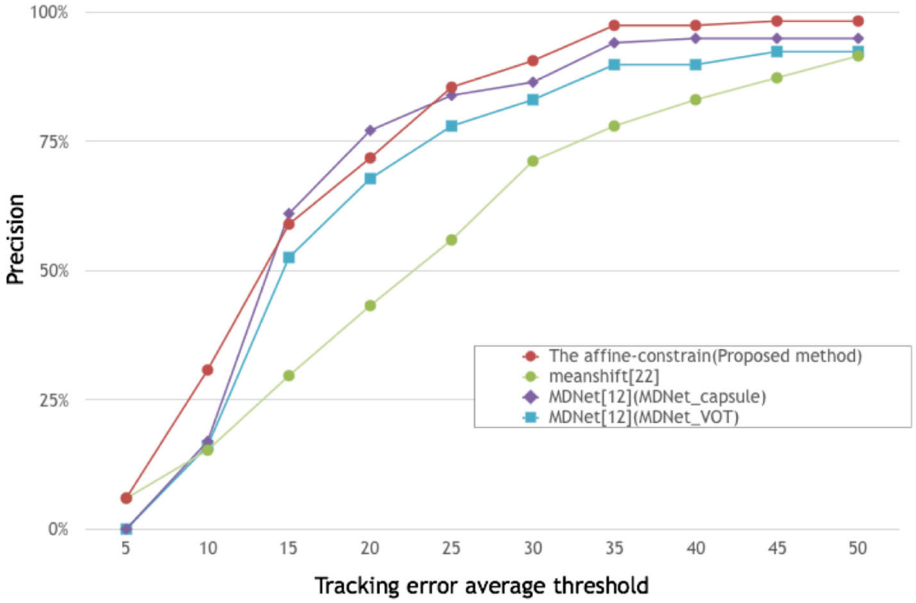


Fig. 8 Precision and success plots per sequence. Here, the performance of the proposed method is compared with that of the mean-shifted method [22] and that of the MDNet method [12] trained using VOT2015 data or capsule endoscopy images

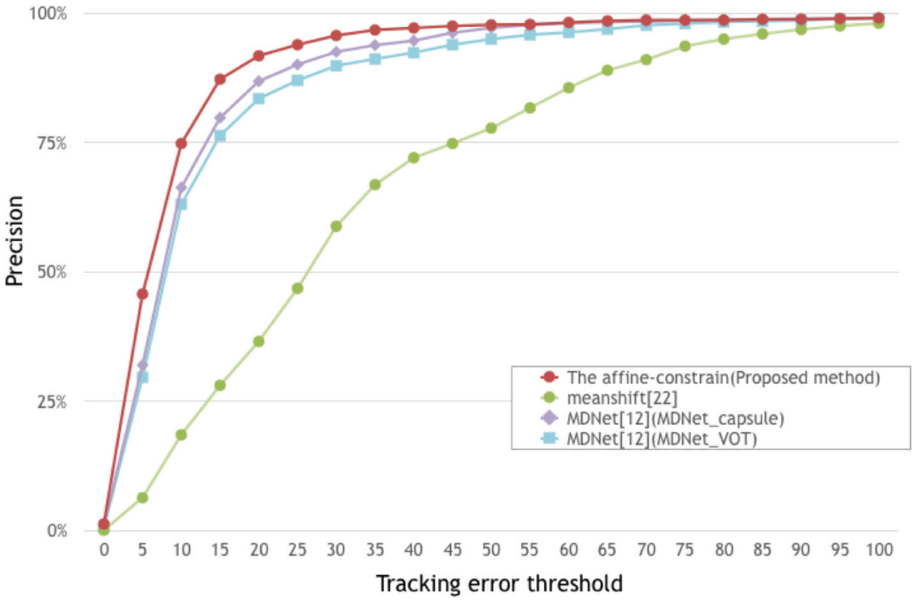


Fig. 9 Precision and success plots per frame, comparing the tracking of our proposed method with that of other state-of-the-art methods

Challenge in 2015 (VOT2015) [14]. The MDNet method requires pre-training of its model before tracking can be accomplished. We therefore prepared two training models. For MDNet_VOT, we use the model that Nam et al. are publishing (https://github.com/HyeonseobNam/MDNet). The model of MDNet_capsule additionally trained the model of MDNet_VOT using about 1000 WCE images.

Our proposed method performed better than both the mean-shifted method and the MDNet method using a VOT2015 training model.

Figure 9 presents the precision and success plots per frame. It clearly illustrates that our proposed method, denoted using the affine constraint, outperforms all other state-of-the-art methods.

Figure 10 demonstrates our ability to track an abnormality, even when there are large movements in the background scene, and when the appearance of the abnormality is difficult to detect with the naked eye. The green rectangle denotes the ground truth region, while the yellow rectangle is our tracking result. Despite irregular movements during the capture of these sequences, the proposed method successfully tracked the abnormality.

Figure 11 provides examples of erroneous tracking. The first row shows a case without a good coarse estimation, and with many similar textures. Thus, in the

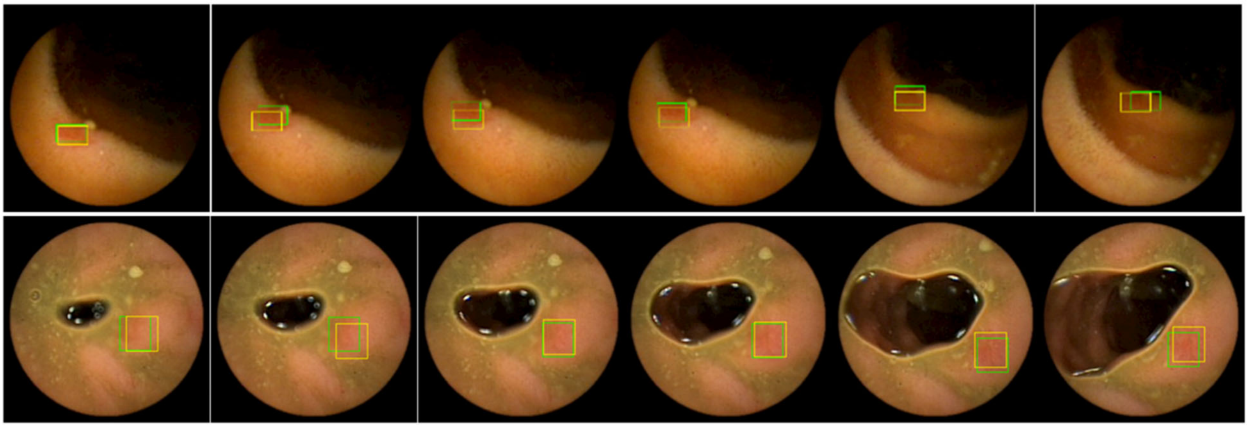


Fig. 10 Examples of the accurate tracking of abnormalities using the affine constraint. The *green rectangles* in each image are the ground truth region, while the *yellow rectangles* are our tracking result

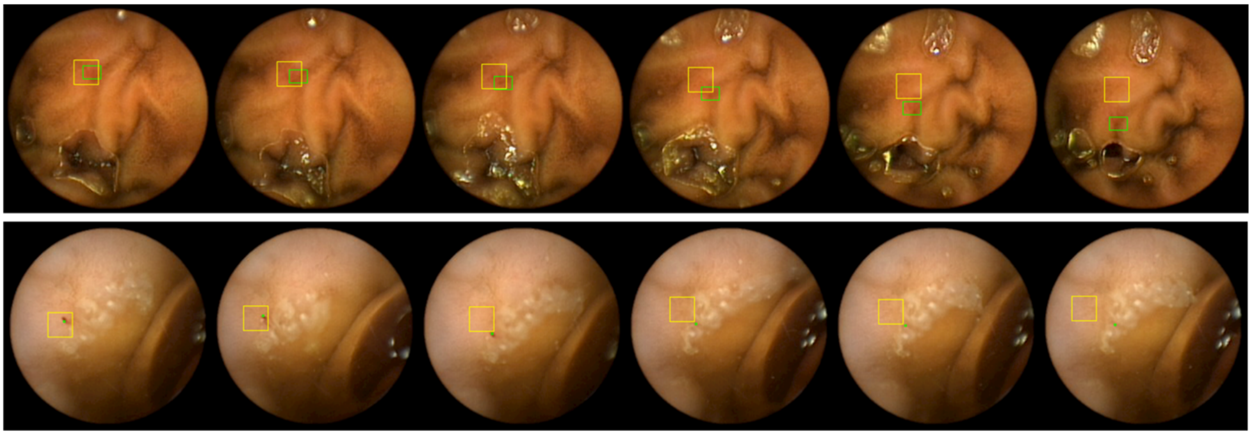


Fig. 11 Examples of erroneous tracking using the affine constraint. *Green rectangles* in each image are the ground truth region, while the *yellow rectangles* show our tracking result

supporter matching step, many incorrect pairs were created. Because the supporter is detected only in the center, the second-row case is an example where the estimation is shifted because the entire image cannot be used. Failure to track the abnormality in first-row cases occurs during determination of the precise position from color information (stage 3) while that in second-row cases occurs during the detection of the supporter.

In a training system, frames in which the abnormality appears and disappears are important. We performed an experiment in which an initial frame, located somewhere within the sequence of images showing an abnormality, is selected manually. Next, the manual tracker searches back and forth to determine the frames in which the abnormality is resolved. Table 1 gives the results of experiments related to detecting the disappearance of the abnormality. "Lost" denotes the case when the abnormality is lost by the tracker, even though it still situated within the video frame. Although the affine-constraint method performs better than other methods, its performance is not satisfactory for training purposes.

4 Conclusions

We proposed a method for tracking areas with abnormalities in an image sequence from capsule endoscopy. It is not unusual that abnormality images in the small intestine do not have strong features, such as sharp edges and distinct color changes. Consequently, the surrounding features are used to estimate the position. The position of the target can be robustly determined from voting relative positions according to an affine transformation of several triplets of supporters. The affine constraint is also effective as shown in the experimental results, which means that the triangular shape is maintained, while still allowing weak deformation between successive frames. In conclusion, the proposed method is able to track an abnormality even if motion displacement is large and the abnormality is indistinguishable. The calculation time was approximately 5 s per frame using a general-purpose personal computer. This calculation time is the same as that for the MDNet method but is later than that of the mean-shifted method. However, we believe that the calculation can be accelerated by multiple-core processing.

Table 1 Performance comparison of our affine-constraint method with angular-constrained and unconstrained tracking methods

	Affine constraint	Angular constraint	No constraint
Correct	114	104	103
Lost	6	16	17

At present, WCE captures images at rates of 2–6 fps, and it is possible to adjust the frame rate with respect to the capsule's speed. Although in our experiments we used images captured at 2 fps, our tracking procedure should also function at 6 fps. Furthermore, our method can easily accommodate variable frame rates, with changing constraints.

We showed that our proposed method is able to track an abnormality, even when motion displacement is large, and its appearance is difficult to detect with the naked eye. As future work, we intend representing color components of both the abnormal area and the adjacent area using a Gaussian mixture model in the fine estimation, to determine the disappearance frame corresponding to the end of the existence of the abnormality.

Abbreviations

CNN: Convolutional neural network; GI: Gastrointestinal; KLT: Kanade-Lucas-Tomasi; MDNet: Multi-domain network; SIFT: Scale-invariant feature transform; SURF: Speeded-up robust features; VOT2015: Visual Tracking Challenge in 2015; WCE: Wireless capsule endoscopy

Authors' contributions

YYan designed the study, analyzed the data, and wrote the initial draft of the manuscript. TE and YYag designed the study and critically reviewed the manuscript. HV conceived the study and participated in its design and coordination and helped draft the manuscript. HO, YF, and TA acquired the data and annotated the abnormality. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]OMRON Corporation, 9-1, Kizugawadai, Kizugawa-shi, Kyoto, Japan. [2]Department of Engineering Informatics, Osaka Electro-Communication University, 18-8 Hatsucho, Neyagawa-shi, Osaka 572-8530, Japan. [3]International Research Institute MICA, Hanoi University of Science and Technology, 1, Dai Co Viet Street, Hanoi, Vietnam. [4]Graduate School of Medicine, Osaka City University, 1-4-3 Asahimachi, Osaka-shi, Osaka 545-8585, Japan. [5]Department of Intelligent Media, The Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki-shi, Osaka 567-0047, Japan.

Received: 4 October 2016 Accepted: 16 February 2017
Published online: 01 March 2017

References

1. Wu CH, Sun YN, Chang CC (2007) Three-dimensional modeling from endoscopic video using geometric constraints via feature positioning. IEEE Trans Biomed Eng 54(7):1199–211
2. Liedlgruber M, Uhl A (2011) A summary of research targeted at computer-aided decision support in endoscopy of the gastrointestinal tract. Technical Report 2011-01. Department of Computer Sciences, University of Salzburg, Austria
3. Berens J, Fisher M, et al (2008) Wireless capsule endoscopy color video segmentation. IEEE Trans Med Imaging 27(12):1769–81
4. Szczypiński P, Klepaczko A, Pazurek M, Daniel P (2014) Texture and color based image segmentation and pathology detection in capsule endoscopy videos. Comput Methods Prog Biomed 113(1):396–411
5. Li B, Meng MQ-H (2009) Texture analysis for ulcer detection in capsule endoscopy images. Image Vis Comput 27(9):1336–42
6. Li B, Meng MQ-H (2009) Computer-aided detection of bleeding regions for capsule endoscopy images. IEEE Trans Biomed Eng 56(4):1032–9
7. Jebarani WSL, Daisy VJ (2013) Assessment of Crohn's disease lesions in wireless capsule endoscopy images using svm based classification. In: ICSIPR. IEEE, Coimbatore. pp 303–7. http://www.ieee.org/conferences_events/conferences/conferencedetails/index.html?Conf_ID=30432
8. Sawant S, Deshpande M (2015) Tumor recognition in wireless capsule endoscopy images. Int J Comput Sci Netw Secur 15(4):85
9. Kalman RE (1960) A new approach to linear filtering and prediction problems. Trans ASME–J Basic Eng 82(Series D):35–45
10. Sorenson HW (1970) Least-squares estimation: from Gauss to Kalman. Ieee Spectr 7(July 1970):63–8
11. Isard M, Blake A (1998) Condensation—conditional density propagation for visual tracking. Int J Comput Vis 29(1):5–28
12. Nam H, Han B (2016) Learning multi-domain convolutional neural networks for visual tracking. In: CVPR. CVPR, IEEE, Washington. pp 4293–302. http://www.ieee.org/conferences_events/conferences/conferencedetails/index.html?Conf_ID=33180
13. Danelljan M, Hager G, Shahbaz Khan F, Felsberg M (2015) Convolutional features for correlation filter based visual tracking. In: ICCV. ICCV, IEEE, Santiago. pp 58–66. https://www.ieee.org/conferences_events/conferences/conferencedetails/index.html?Conf_ID=33071
14. Kristan M, Matas J, Leonardis A, Felsberg M, Čehovin L, Fernandez G, Vojir T, Häger G, Nebehay G, Pflugfelder R, Gupta A, Bibi A, Lukežič A, Garcia-Martin A, Saffari A, Petrosino A, Montero AS, Varfolomieiev A, Baskurt A, Zhao B, Ghanem B, Martinez B, Lee B, Han B, Wang C, Garcia C, Zhang C, Schmid C, Tao D, Kim D, Huang D, Prokhorov D, Du D, Yeung DY, Ribeiro E, Khan FS, Porikli F, Bunyak F, Zhu G, Seetharaman G, Kieritz H, Yau HT, Li H, Qi H, Bischof H, Possegger H, Lee H, Nam H, Bogun I, Jeong J-c, Cho J-i, Lee JY, Zhu J, Shi J, Li J, Jia J, Feng J, Gao J, Choi JY, Kim JW, Lang J, Martinez JM, Choi J, Xing J, Xue K, Palaniappan K, Lebeda K, Alahari K, Gao K, Yun K, Wong KH, Luo L, Ma L, Ke L, Wen L, Bertinetto L, Pootschi M, Maresca M, Danelljan M, Wen M, Zhang M, Arens M, Valstar M, Tang M, Chang MC, Khan MH, Fan N, Wang N, Miksik O, Torr PHS, Wang Q, Martin-Nieto R, Pelapur R, Bowden R, Laganiere R, Moujtahid S, Hare S, Hadfield S, Lyu S, Li S, Zhu SC, Becker S, Duffner S, Hicks SL, Golodetz S, Choi S, Wu T, Mauthner T, Pridmore T, Hu W, Hübner W, Wang X, Li X, Shi X, Zhao X, Mei X, Shizeng Y, Hua Y, Li Y, Lu Y, Li Y, Chen Z, Huang Z, Chen Z, Zhang Z, He Z (2015) The Visual Object Tracking VOT2015 challenge results. In: Visual Object Tracking Workshop 2015 at ICCV2015. Computer Vision Workshop (ICCVW), 2015 IEEE International Conference on, Santiago. pp 564–86
15. Kanazawa Y, Uemura K (2006) Wide baseline matching using triplet vector descriptor. In: BMVC. BMVC, Edinburgh. pp 267–276. http://www.bmva.org/bmvc/?id=bmvc
16. Grabner H, Matas J, Gool LJV, Cattin PC (2010) Tracking the invisible: learning where the object might be. In: CVPR. CVPR, IEEE, San Francisco. pp 1285–1292
17. Yanagawa Y, Echigo T, Vu H, Okazaki H, Fujiwara Y, Arakawa T, Yagi Y (2012) Tracking abnormalities in video capsule endoscopy using surrounding features with a triangular constraint. In: ISBI. ISBI, IEEE, Barcelona. pp 578–581. https://www.ieee.org/conferences_events/conferences/conferencedetails/index.html?Conf_ID=17944
18. Tomasi C, Kanade T (1991) Detection and tracking of point features. Technical report. International Journal of Computer Vision. http://link.springer.com/journal/11263
19. Lowe DG (1999) Object recognition from local scale-invariant features. In: ICCV. ICCV, IEEE, Kerkyra. p 1150
20. Bay H, Tuytelaars T, Gool LV (2006) SURF: speeded up robust features. In: ECCV. ECCV, Springer, Graz. pp 404–17. http://www.springer.com/la/book/9783540338321
21. Vu H, Echigo T, Yagi K, Okazaki H, Fujiwara Y, Yagi Y, Arakawa T (2011) Image-enhanced capsule endoscopy preserving the original color tones. In: in Proceeding of Workshop on Computational and Clinical Applications in Abdominal Imaging 2011, LNCS 6668 (In Print). Springer Berlin Heidelberg, Toronto. http://link.springer.com/chapter/10.1007/978-3-642-28557-8_5
22. Collins RT (2003) Mean-shift blob tracking through scale space. In: CVPR. CVPR, IEEE, Madison. pp 234–40