Realtime computer-aided object detection in endoscopic screening

Katsiaryna Halavataya
Intelligent systems dept.
Faculty of Radiophysics and Computer Technologies
Belarussian State University
Minsk, Belarus
katerina-golovataya@yandex.ru

Aliaksandr Kurachkin
Intelligent systems dept.
Faculty of Radiophysics and Computer Technologies
Belarussian State University
Minsk, Belarus
lawliet29@gmail.com

***Abstract*—The project focuses on using computer vision techniques to provide visual support and highlighting when performing endoscopic screening. The system is meant to provide visual clues to the physician performing the screening, highlighting regions of interest and polyps in real time, in order to increase the evaluation accuracy of esophageal, gastric and colon cancers.**

***Keywords*—medical image processing, computer vision, realtime object detection, machine learning**

I. Introduction

Endoscopy is a widespread approach in medical diagnostics that involves a series of procedures for getting visual information from inside the human body in order to examine internal organs. The defining triat of all endoscopic research is the acquisition of colored realistic representative image, image sequence or video as part of the procedure, which greatly helps in early detection, diagnosis and treatment of a wide array of diseases. Most of the modern videoendoscopic system presume the acquisition of high-definition high-framerate video in realtime from the endoscopic camera inside the human body, equipped with additional light source, and allow to control the examination in order to acquire more information about specific areas that might require the physician's attention.

The effectiveness of endoscopic examination is determined by a multitude of factors - patient's preparation for the procedure, equipment quality, physician's skills in working the equipment, ability to spot areas that require further attention and ability to make educated decisions based on the data acquired, sometimes during the procedure in order to control its further flow. Additional equipment features like retroflexion, second view, water/air inflation, etc. also affect the effectiveness of the procedure. However, modern research indicates that physician's own skills play a great role, which led to several medical care institutions of the USA, including U.S. Multisociety Task Force on Colorectal Cancer, American College of Gastroenterology, American Society for Gastrointestinal Endoscopy Task Force on Quality, to propose in 2002 a number of research quality indicators to evaluate physician's effectiveness during the research. In following years, most of these indicators were adopted in other countries, and are nowadays considered a worldwide standard by various health care institutions throught the world.

One of the most important quality indicators of endoscopy screening quality is so-called personal adenoma detection rate, or ADR. A physician's ADR is the proportion of individuals 50 or more years of age undergoing a complete screening colonoscopy who have one or more adenomas, or polyps, detected. A typical value for good ADR is at least 15% for women and 25% for men. It should be noted that while not all detected adenomas may be potentially dangerous, it's still important to spot them during the screening in order to obtain enough information to make sure it poses no threat to the patient. [1]

The paper focuses on development of automated object detection system to aid with screening process. The main constraint on the system is the ability to work in real time – object detection must be performed during the procedure in order to give the physician necessary time to decide if the detected object should be examined further and to minimize overall examination time.

II. Realtime object detection techniques

The problem of detecting and localizing objects on the image is a well-known one in the field of computer vision. The basic definition of the problem can be postulated as follows: given an image that can potentially contain an object of interest, detect said object by making sure that it conforms to specific kinds of visual, spatial, pattern and brightness criteria; produce, if possible, a bounding area (usually bounding rectangle) containing that object; and, finally, produce, if possible, a collection of pixels that belong to the detected object. [2], [3]

The simplest form of object detection may be defined as a binary classifier with image dimensions as bounding rectangle. Algorithm implementation based on this definition provides no information about the actual location of detected object on the image, since the bounding rectangle covers every point that belongs to the image itself. The resulting classes of this binary classification can be treated as "object present on the image" and "object absent on the image" for positive and negative classification, respectively. Most of the times, binary classification output is smoothed, and classification result x is represented as a continious value over the range of $x \in [0; 1]$. With this definition, classification result can be treated as a confidence score of an object being present on the image.

Binary classifiers can be used as a base for creating multiclass classifiers – given n independent binary classifiers that produce $x_i \in [0; 1]$ confidence score of an object of i-th type being present on the image, it's possible to construct a multiclass n-dimensional vector $\vec{x} = (x_1, x_2, ..., x_n)$ with confidence scores for each of the n classes. Choosing a "winning" class or classes can be then implemented as choosing classes with maximum confidence scores, or choosing classes based on a certain confidence score threshold. [2], [4], [5]

One step further is the implementation of actual localization of detected objects on the image. Usually localization presumes finding a rectangular region on the image that encompasses the target object of the classification. It's possible to implement basic localization using a given binary image classifier with a simple brute force approach by iterating over a fixed number of blocks and their adjacency combinations and running classifier with each block to find the region that produces the highest confidence value.

After the actual localization, sometimes it's also necessary to produce the actual uneven boundaries of detected object, i.e. perform classification on a pixel-per-pixel basis in order to find the exact pixels that comprise the classified object on the image. [6], [7]

The most common approach to real-time object detection presumes the usage of features. Features of the image are usually defined as points of the image that represent its characteristic visual triat in a form of some well-known visual abstraction. There is no general criteria that makes any given pixel or any given region of the image a feature – these are usually defined depending on the specific problem and application; however, most commonly feature points are chosen based on sharp brightness difference, thus corresponding to edges and corners of a given image.

The algorithms for computing a set of feature points (or keypoints) of the image are called feature or keypoint detectors. These include most common edge detectors, corner detectors and blob detectors. Some higher-order feature detectors aim to produce keypoints that are scale- and transform-invariant, i.e. don't depend on a relative scale, rotation and skew of their spatial surrounding area. [2], [4], [8]

An actual classifier based on features also requires some notion of feature correspondence or feature comparison. The main idea is that the sought object on an image posesses some features that are similar – to a certain degree – to the set of reference features of a model object that is being detected. Comparison of features is usually performed in a feature metric vector space, and projections from keypoint to feature space are performed by algorithms called feature descriptors.

A feature descriptor f is a projection of any point p_{ij}^I of any image I to an n-dimensional metric vector space F:

$$f(p_{ij}^I) = \vec{v}_{ij}^I \in F \tag{1}$$

Since vector space F is also a metric space, an appropriate metric is defined on it:

$$m : F \times F \to \mathbb{R} \tag{2}$$

Two arbitary points, p^{I_1} of image I_1 and p^{I_2} of image I_2, are considered similar by a feature descriptor (1) if their feature vectors $\vec{v}_1 = f(p^{I_1})$ and $\vec{v}_2 = f(p^{I_2})$ are similar by measure of metric (2), or $m(\vec{v}_1, \vec{v}_2) < t$. The threshold t is selected based on a specific descriptor implementation.

Histogram-based descriptors use histogram analysis methods to produce feature vectors – the subregions of the immediate keypoint surroundings are aggregated to magnitude and orientation values and sampled across a fixed-size grid into histogram bins, and the descriptor itself is defined as all values of those histograms. Feature vectors of histogram-based methods usually have a very high number of dimensions (e.g. 128 for SIFT), and the most common distance metric (euclidean multidimensional vector distance) can be computationally complex to calculate.

Subfeature-based descriptors consist of multiple subfeature detectors, each producing its single distinct output value based on sampling pattern, which is then usually normalized. The most typical sampling patterns produce pair-by-pair comparison between specific points of surrounding region. More complex subfeature detectors are also coupled with orientation compensation pattern to reduce the effect of affine transform on pair-by-pair comparison of sampling pattern. Each subfeature value is then stored in the resulting feature vector. Since each subfeature is distinct, they can only be compared by their corresponding values, so the most common distance metric is the sum of differences for the same features.

The most repetative and computationally expensive task when working with feature descriptors is usually the actual feature vector comparison, i.e. the calculation of a metric m (2). Most of the time the algorithmic complexity of feature descriptors is high enough as it is; moreover, common recognition tasks presume comparison of two sets of feature points by performing comparison for every possible pair. In this context, of particular interest are subfeature-based feature extractors for which each of the subfeatures can be normalized and then binarized with an appropriate threshold. That way, each subfeature output becomes a binary value, and feature vector space becomes an n-dimensional boolean $\mathbb{B}^n$. The most common metric for these descriptors becomes the Hamming distance – a count of non-matching subfeatures. Descriptors that produce binary feature vectors are called binary descriptors, and are most widely used in real-time processing tasks, because calculating Hamming distance between to feature vectors is extremely fast.

The classical approach to object detection was proposed by P. Viola in 2001 paper "Rapid Object Detection using a Boosted Cascade of Simple Features". It is a machine learning approach that uses Haar-like feature extractors and appropriate feature descriptors to match around 6000 features in a 24x24 pixel windows, combined in a cascade of classifiers. This approach is still widely used for face recognition problems and is computationally effective enough to be used in real-time. Various modifications of this approach were proposed in the following years, making it more suitable for specific pattern recognition tasks. [9]

Another, more recent approach to object detection uses region-based active contour models for image segmentation. The boundaries are computed and adjusted iteratively using edge-based and region-based terms of particular feature extractor and descriptor combination, defined as an optimization problem of a local cost function of boundary inconsistency accros feature space.

A more modern approach to object detection and localization is the usage of convolutional neural networks. Convolutional neural networks differ from more traditional object detection methods because the network itself is trained to produce and accentuate features it requires for a particular object classification. By examining the output of the deeper layer neurons it's also possible to evaluate the boundaries of a given object. While training a convolutional neural network usually takes a significant amount of time, fully trained networks are actually performant enough to perform object classification and detection in real-time.

III. Real-time polyp detection

There are multiple specifics that must be taken into account when working with videoendoscopic images and image sequences.

First, most of the videoendoscopes come equipped with a wide-angle lens camera. The non-linear transformation caused by a wide-angle lens produces the effect most commonly known as fisheye distortion. Fisheye lenses capture the light of not only immediate forward frame area, but also of objects reflected from around its vicinity. Because of the nature of such a distortion, objects closer to the edges of the image appear much larger than they actually are, and objects closer to the center appear smaller. Also, straight lines moving away from the screen towards the center gain a curvature that may be recalculated given known lens parameters. [8]

For most practical applications of digital image processing distortion correction is a necessary pre-processing step. It allows to eliminate the inconsistencies of scale and curvature which are determinal when using keypoint-based feature extractors and object detection algorithms.

Of particular note is the narrow-band imaging (NBI) technique implemented in most of the modern videoendoscopic systems. The main idea of NBI is applying a set of color filters on wavelengths that correspond to typical color chroma of blood, blood vessels, background tissue and other objects most likely present on endoscopic image, and then normalizing the remaining wavelength range non-linearly. The result is usually much more visually contrast image that makes expert evaluation of the region much easies.

All of the above produce a set of requirements that an automated real-time polyp detection system must meet:

- Detection must be invariant against image rotation. This requirement is based on the fact that a certain region may be observed under different angles, and its detection and classification must keep working in those cases.
- Detection must be distortion-invariant. A specific pattern of polyp should be detected regardless of the way it's distorted, i.e. on distorted image detector must produce accurate results on the edges and towards the center of the image.
- Detection should work regardless of relative contrast and absolute color values of the objects. Since regular image, post-processed image and NBI image produce different color representations of the same region, detector must use spatial and differential features in order to classify objects.
- Detection visualization should not interfere with the procedure. Detected object boundaries should not obstruct important parts of the image, while at the same time location of detected object should be clear and distinct enough to be able to spot it easily.
- Detection, as mentioned several times in this paper, should be performed in real time. This requirement means that detection should work on a separate framerate than the main camera in order to prevent input lag, and it should be fast enough to have a detection performance of 15-20 frames per second to be able to keep up with the main screen.
- Detection should be precise. False-negative detections (i.e. failing to detect an object) mean that a potentially dangerous polyp can be overlooked, while false-positive detections (i.e. detecting an object where there is none) may unnecesarily divert physician's attention, thus complicating and prolonging the procedure.

Image rotation invariance can be achieved by using scale-invariant feature extractors. Most of the modern feature extractors already provide scale-invariance i.e. descriptors of features remain similar when the image is subjected to simple affine transformations. For the actual implementation of the system, it was decided to use ORB (Oriented FAST and Rotated BRIEF) feature detector and extractor. It provides numerous significant advantages over more traditional Scale-Invariant Feature Transform based approaches in a significant increase of efficiency; moreover, descriptors produced by ORB are binary, which means that keypoint matching is extremely fast. [4]

Distortion invariance is implemented using a simple distortion correction algorithm that uses standard fisheye distortion model with adjusted parameters acquired using endoscopic camera calibration. While the actual transformation of the entire image can be computationally expensive, actual per-pixel transformation can be postponed until its evaluation is required. [8], [10] Moreover, keypoint detection for some of the video frames can be skipped entirely and instead localized to areas around keypoints detected on the previous frame, thus the evaluation of spatial brightness and distortion correction transformation will only be required in these areas. The system re-calculates a new set of keypoints on each 20th frame.

Color invariance is achieved by simplifying input frame colorspace. There are 2 modes supported – using normalized blue color component of RGB image representation and using normalized brightness component of HSB image representation. Normalization makes sure that relative contrast doesn't affect the results as much.

Unobtrusiveness of detection result visualization is hard to achieve. For initial implementation, it was decided to use a rectangle selection to outline the bounding frame of detected object with the ability to hide detection markers at will if they happen to interfere with the observation.Example frame with outlined detected object is presented on Figure 1.

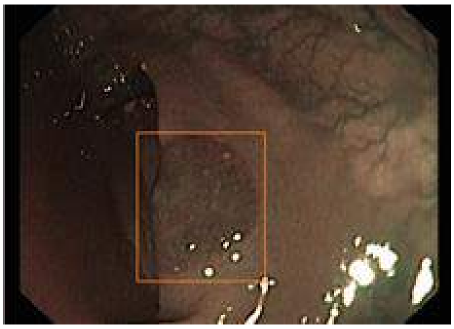

Figure 1. Example of a figure caption.

Finally, in order to optimize the precision of detection it is possible to adjust the sensitivity of the resulting detection. In order to match the sought objects, a comprehensive feature vector sets were built by using a mapped image sets built from several colonoscopy screening videos containing known types of polyps that should be detected by the system, evaluated by an expert. While the precise sensitivity can only be determined on a case-by-case basis, a reasonable default value is provided. Also, the system optionally provides suggestions to enable or disable certain endoscopy system built-in image enhancements when output confidence score is not high enough to qualify as exact match.

IV. Conclusion

Implemented software complex can be used as a decision support system for endoscopic examination. The system is able to detect suspicious objects in real-time during the screening, make them visually clear to the physician performing the procedure in order to optimize the procedure time, emphasise the attention on certain automatically detected areas during the screening procedure which, in turn, should serve to increase the effective ADR of the physician.

Future system enhancements include the implementation of multi-class detection in order to not only spot suspicious objects on the image, but to also provide insights about the exact type of the object. This can further be expanded to propose suggestions about the optimal diagnosis and even further treatment. Also, these tasks can be also performed after the initial examination by analyzing the resulting video, thus lifting the real-time processing constraint. Since object detection relies heavily on calculating and evaluating keypoints using feature detectors and descriptors, it's also possible to include different types of processing that makes use of feature extraction. One of those is the problem of 3-dimensional spatial reconstruction of the scenes present on the video. This can prove benificial in a more detailed analysis of areas of interest.

References

[1] D. A. Corley et al, *Adenoma Detection Rate and Risk of Colorectal Cancer and Death*. The New England Journal of Medicine; vol. 370, pp. 1298-1306; 2014

[2] L. G. Shapiro, G .C. Stockman, *Computer Vision*. New Jersey, Prentice-Hall, 608 p.; 2001

[3] C. Barata, J. S. Marques, J. Rozeira, *The role of keypoint sampling on the classification of melanomas in dermoscopy images using bag-of-features*. Iberian Conference on Pattern Recognition and Image Analysis, pp. 715-723. Springer, Berlin, Heidelberg, 2013.

[4] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, *ORB: an efficient alternative to SIFT or SURF*. 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2564-2571, IEEE, 2011

[5] K. M. Yi, E. Trulls, V. Lepetit, P. Fua, *Lift: Learned invariant feature transform*. European Conference on Computer Vision, pp. 467-483. Springer International Publishing, 2016.

[6] W. Wu, A. Y. C. Chen, L. Zhao, J. J. Corso, *Brain Tumor detection and segmentation in a CRF framework with pixel-pairwise affinity and super pixel-level features*. International Journal of Computer Aided Radiology and Surgery; vol. 9, pp. 241-253; 2014

[7] M. Gabryel, M. Korytkowski, R. Scherer, L. Rutkowski, *Object detection by simple fuzzy classifiers generated by boosting*. International Conference on Artificial Intelligence and Soft Computing, pp. 540-547. Springer, Berlin, Heidelberg, 2013.

[8] A. Sotiras, C. Davatzikos, N. Paragios, *Deformable Medical Image Registration: A Survey*. IEEE Transactions on Medical Imaging, vol. 32, issue 7, pp.1153-1190, IEEE, 2013

[9] P. Viola, M. Jones, *Rapid object detection using a boosted cascade of simple features*. IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1., pp. 511-519, IEEE, 2001

[10] A. Foncubierta-Rodriguez, H. Müller, A. Depeursinge, *Region based volumetric medical image retrieval*. SPIE Proceedings, vol. 8674, p. 10. 2013.

АВТОМАТИЗИРОВАННОЕ ДЕТЕКТИРОВАНИЕ ОБЪЕКТОВ В РЕАЛЬНОМ ВРЕМЕНИ ПРИ ПРОВЕДЕНИИ ЭНДОСКОПИЧЕСКОГО СКРИНИНГА

Головатая Е.А., Белорусский Государственный университет

Курочкин А.В., Белорусский Государственный университет

В данной работе рассматриваются методы компьютерного зрения для предоставления визуальной поддержки и выделения объектов при проведении процедуры эндоскопического скрининга. Разработанная авторами система предоставляет визуальные подсказки специалисту, осуществляющему скрининг, и выделяет образования и другие области интереса в реальном времени, с целью повышения точности диагностики различных видов рака.