

# Interactive Labeling of WCE Images

Michal Drozdal<sup>1,2</sup>, Santi Seguí<sup>1,2</sup>, Carolina Malagelada<sup>3</sup>,  
Fernando Azpiroz<sup>3</sup>, Jordi Vitrià<sup>1,2</sup>, and Petia Radeva<sup>1,2</sup>

<sup>1</sup> Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra, Spain

<sup>2</sup> Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Barcelona, Spain

<sup>3</sup> Hospital de Vall d'Hebron, Barcelona, Spain

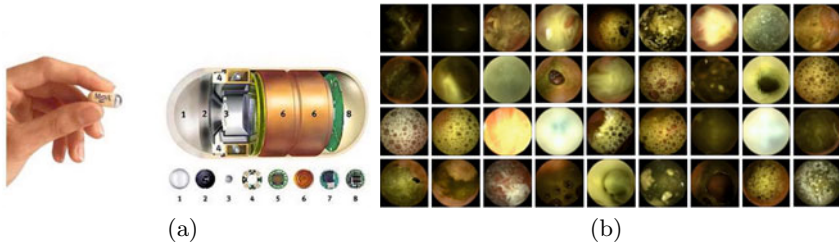
**Abstract.** A high quality labeled training set is necessary for any supervised machine learning algorithm. Labeling of the data can be a very expensive process, specially while dealing with data of high variability and complexity. A good example of such data are the videos from Wireless Capsule Endoscopy. Building a representative WCE data set means many videos to be labeled by an expert. The problem that occurs is the data diversity, in the space of the features, from different WCE studies. That means that when new data arrives it is highly probable that it will not be represented in the training set, thus getting a high probability of performing an error when applying machine learning schemes. In this paper an interactive labeling scheme that allows reducing expert effort in the labeling process is presented. It is shown that the number of human interventions can be significantly reduced. The proposed system allows the annotation of informative/non-informative frames of the WCE video with less than 100 clicks.

**Keywords:** WCE, interactive labeling, online learning, LSH.

## 1 Introduction

For many machine learning applications, the compilation of a complete training set for building a realistic model of a given class of samples is not an easy task. The motto “*there is no data like more data*” suggests that the best strategy for building this training set is to collect as much data as possible. But in some cases, when dealing with problems of high complexity and variability, the size of this data set can grow very rapidly, making the learning process tedious and time consuming. Time is expended in two different processes: 1) the labeling process, which generally needs human intervention, and 2) the training process, which in some cases exponentially increments computational resources as more data is obtained.

If we consider training as a sequential (in time) process, the training problem can be partially overcome by the integration of online learning methodologies and an “intelligent” reduction of the samples to be added into the training set. Given a training set at a given time, an “intelligent” system should add to the training set only those data samples that were not represented, or under-represented, in the previous version of the set. In other words, the training set



**Fig. 1.** (a) The wireless video capsule; (b) Non informative frames

should be enlarged by those new data samples that enrich the representability of the classification models while avoiding unnecessary sample redundancy.

But even in the former situation, the labeling process is still a problem. To overcome this problem, we should use a labeling process which minimizes the number of human interventions on the new samples. One possible strategy is to embed a model of the data into a sequential labeling process, for proposing to the human operator a given label for every sample. Then, the human operator faces two possible decisions: to accept the model proposal or to change the label of the sample. In practice, these two choices have a non symmetric cost for the human operator: accepting the model proposal can be efficiently implemented with a low cognitive load for the operator, while changing a label has a larger cost. This effort can be measured by the number of “clicks” the operator performs.

Wireless Capsule Endoscopy (WCE) image analysis (see Fig. 1(a)) is a clear scenario where these problems arise. The WCE contains a camera and a full electronic set which allows the radio frequency emission of a video movie in real time. This video, showing the whole trip of the pill along the intestinal tract, is stored into an external device which is carried by the patient. These videos can have duration from 1h to 8h, what means that the capsule captures a total of 7.200 to 60.000 images. WCE videos have been used in the framework of computer-aided systems to differentiate diverse parts of the intestinal tract[4], to measure several intestinal disfunctions [11] and to detect different organic lesions (such as polyps [6], bleeding [5] or general pathologies [2]). In most of these applications, machine learning plays a central role to define robust methods for frame classification and pattern detection.

A common stage in all these research lines is the discrimination of informative frames from non-informative frames. Non-informative frames are defined as frames where the field of view is occluded. Mainly, the occlusion is caused by the presence of intestinal content, such as food in digestion, intestinal juices or bubbles (see Fig. 1(b)). The ability of finding non-informative frames is important since: 1) generally, it helps to reduce time of video analysis, and 2) since the majority of non-informative frames are frames with intestinal content which information can be used as an indicator for intestinal disfunctions [8].

The main strategy in the search of non-informative frames is the application of machine learning techniques in order to build a two-class classifier. Generally, non-informative frames are characterized by their color information [1]. Robust

classifiers can be built when the training set is representative of the data population. The problem that occurs when dealing with WCE videos is the high color variability of non-informative frames in different videos. Therefore, it is probable that, as new videos become available, a new video with significantly different intestinal content color distribution will be added to the training set. A naive approach for the labeling process of this video could mean the manual annotation of up to 50.000 video frames.

The main contribution of this paper is a method for interactive labeling that is designed to optimally reduce the number of human interventions during the labeling process of a new video. The process is applied to the labeling of non-informative frames for WCE image analysis. In the active learning algorithm the key idea is that machine learning algorithm can achieve greater accuracy with fewer labeled training instances if it is allowed to choose the data from which it learns [10,7]. However, the goal of our system is basically to enlarge the training set while minimizing human intervention, not to minimize the overall classification error. This is done by finding an optimal classifier adaptation scheme for non-represented frames in the original training set.

The rest of the paper is organized as follows: Section 2 introduces the interactive labeling method, Section 3 describes the implemented interactive labeling system for WCE images. In Section 4, we present experimental results, and finally, in Section 5 we expose some conclusions and remarks on future research.

## 2 Interactive Labeling

Our purpose is to design an interactive labeling system that should allow, in an efficient way, 1) to detect frames that are not represented in the training set, 2) to obtain statistics of informative and non-informative frames that are not represented in the training set, and 3) to be able to iteratively increase the representability of the training set in an “intelligent” way by reducing significantly the number of clicks relates to manual labeling.

To this aim, we propose the following algorithm for interactive labeling of a set of new images optimizing the user feedback (see Alg. 1).

The critical step in order to minimize the number of clicks is to choose a good criterion for Step 6, since it represents the main strategy of choosing the order of presentation of the samples to be labeled by the user.

To this end, we studied three different sorting policies for the elements of  $N$ . These policies are based on the following criteria: 1) to choose those elements that are far from the training data  $L$  and far from the boundary defined by  $M_i$ , 2) to choose those elements that belong to the most dense regions of  $N$ , and 3) to choose the elements in a random way.

More specifically, we define them in the following way:

**Criterion 1 (C1):** *Distance of data to the classifier boundary and training data.* In this criterion two measurements are combined: 1) The data are sorted from the farthest to the nearest distance with respect to the classifier boundary. This scheme assumes that the classifier, while proposing labels, will commit

---

**Algorithm 1.** Interactive labeling algorithm.

---

```

1: Let be  $L$  a set of labeled data,  $M_1$  a discriminative model trained on this set,  $U$  a
   set of all unlabeled data samples from a new video and  $C$  a criterion to select the most
   informative frames from  $U$ ;
2: Select the subset of samples  $N = \{x_j^N\}$  from  $U$  such that they are considered as
   under-represented by the labeled set  $L$ .
3: Evaluate the subset  $N$  with  $M_1$ , assigning a label  $l_j$  to every data sample  $x_j$  from  $N$ .
4:  $i=1$ ;
5: while there are elements in  $N$  do
6:   Evaluate the elements of  $N$  with respect to the criterion  $C$  and get the set of  $n$  most
   informative samples  $I \subset N$  (with the purpose of minimizing the expected number of
   expert clicks).
7:   Delete the elements of  $I$  from  $N$ .
8:   Present the samples from  $I$  to the user with their associated label.
9:   Get the user feedback (nothing for samples with correct labels, one click for each
   wrongly classified sample).
10:  Update  $L$  by adding the elements of  $I$  and its user-corrected label.
11:  Perform an online training step for  $M_i$  by adding the elements of  $I$  to the model,
   getting  $M_{i+1}$ .
12:  Evaluate the elements of  $N$  with  $M_{i+1}$ , assigning a label  $l_j$  to every data sample  $x_j$ 
   from  $N$ .
13:   $i = i + 1$ ;
14: end while

```

---

errors with higher probability for the samples that are far from the boundary than for the data that are relatively close to boundary. 2) The data are sorted from the farthest to the nearest with respect to the training data. This scheme assumes that the classifier, while proposing labels, will commit errors with higher probability for the samples that are far from the training set than for the data that are relatively close to the known data. A final sorting is performed in the data by adding the ranking indices of the two previously described schemes.

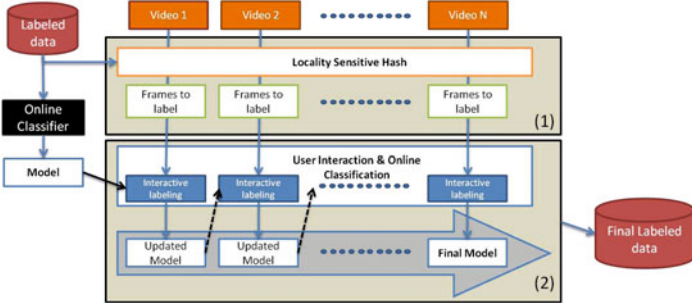
**Criterion 2 (C2): Data density.** Each sample is sorted decreasingly with respect to a data density measure in its environments. This scheme assumes that the classifier should learn more quickly if we first provide samples from the zones with higher density. Data density can easily be computed as the mean distance to the  $k$ -nearest neighbors of the sample.

**Criterion 3 (C3): Random order.** The order of presentation of the samples randomly determined.

### 3 The Interactive Labeling System

The goal of the interactive labeling system is two-fold: 1) to detect, for each new video, the set of frames that are not represented in the training set, and 2) to label those frames with minimal user effort. To this end, we propose a system design with two main components(see Fig. 3):

1. A data density estimation method that allows fast local estimation of the density and distance of a data sample to other examples, e.g. from the training set (see Step 3 of the algorithm for interactive labeling).
2. An online discriminative classifier which allows to sequentially update the classification model  $M_i$  of thousands of samples (see Step 2, 10 and 11 of the algorithm for interactive labeling).



**Fig. 2.** The interactive labeling system architecture with its two main components: 1) Detection of frames not represented in the training set and 2) Labeling of frames and model enlarging using online classifier method

### 3.1 Fast Density Estimation

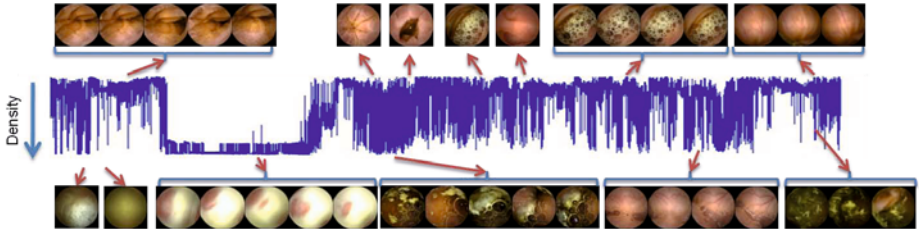
As previously commented, the local density of a data sample  $x_i$  with respect to a data set  $C$  can be easily estimated by computing the mean distance from  $x_i$  to its  $k$ -nearest neighbors in  $C$ . The simplest solution to this problem is to compute the distance from the sample  $x_i$  to every sample in  $C$ , keeping track of the “ $k$ -best so far”. Note that this algorithm has a running time of  $O(nd)$  where  $n$  is the cardinality of  $C$  and  $d$  is the dimensionality of samples.

Because of excessive computational complexity of this method for large data sets, we need a flexible method that allows from one side, effective measurements of characteristics of large data and, from the other side, introducing new unseen data into the training set for enlarging the data representation. An example of such flexible method is Locality Sensitive Hashing[3]. LSH allows to quickly find a similar sample in a large data set. The basic idea of the method is to insert similar samples into a bucket of a hash table. As each hash table is created using random projections over the space, several tables can be used to ensure an optimal result [3]. Another advantage of LSH is the ability to measure the density of data in given space vicinity by analyzing the number of samples inside the buckets (Fig. 3.1). In order to evaluate if the new sample improves the representation of the data set, the space density of the training set is estimated. If the new sample is in a dense part of the space then the sample is considered redundant and thus it is not considered in order to improve the training set. Otherwise, the sample is used to enlarge the training set.

The density  $D$  of the sample  $x$  is estimated according to the formula:

$$D(x, T_r) = \sum_{i=1}^M ||B_i|| \quad (1)$$

where  $M$  is the number of hash tables,  $||B_i||$  is the number of elements in the bucket where the new element  $x$  is assigned and  $T_r$  represents the training set.



**Fig. 3.** Example of training set density estimation for a test video using LSH. The images show the zones of high and low density with respect to given labeled set  $L$ .

The subset of not-represented samples in the training set  $N = \{x_1^*, \dots, x_m^*\}$  from new unlabeled data  $U$  is defined as:

$$N = \{\forall x \in U : D(x, T_r) < T\} \quad (2)$$

where  $T$  represents a fixed threshold.

Note that (2) expresses the condition that the new samples fall in buckets with low density of the training set. That is if the new sample is in a dense part of the space then the sample is not considered to improve the training set. Otherwise, the sample is used to enlarge the training set.

### 3.2 Online Classifier

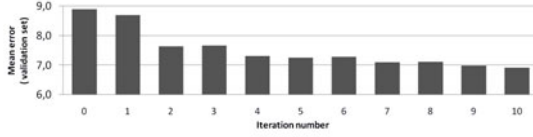
Taking into account that our classifier must be retrained with thousands of images/feature vectors of up to 256 components, using an online classifier is a must. Online classifiers are able to update the model in a sequential way, so the classifier, if needed, can constantly learn from new data, improving the quality of label proposal process. In order to optimize the learning process, the data are sorted according to the previously described criteria. A kernel-based online Perceptron classifier [9] is used because of its simplicity and efficiency. As previously mentioned, the main information used to detect non-informative frames is the color. In order to reduce the dimensionality of the data, each image represented as 24 million colors is quantized into 256 colors. As a result, each frame is represented by 256 color histogram. The score, for a given sample, takes this form:

$$S(x) = \sum_{j=1}^K \alpha_j K(v_j, x) \quad (3)$$

where  $\langle (v_1, \alpha_1), \dots, (v_k, \alpha_k) \rangle$  are the set of training vectors with the corresponding real estimated weights  $(\alpha_1, \dots, \alpha_k)$  by the learning algorithm when minimizing the cumulative hinge-loss suffered over a sequence of examples and  $K()$  is a kernel function (in our case, we apply Radial Basis Function).

## 4 Results

For our experiments, we considered a set of 40 videos obtained using the WCE device. 10 videos were used to build the initial classification model  $M_1$ , and the



**Fig. 4.** Mean error on validation set of 20 WCE videos

other 10 to evaluate the proposed interactive labeling system. In the test, the 10 videos were sequentially processed. If needed, at each iteration, the training set could be increased by a new set of frames that improves the data representation. Additionally the validation set of 20 videos were used in order to evaluate the error of the final informative/non-informative frames classifier.

In the experiments we show that: 1) the proposed system reduces the effort needed for data labeling, 2) the first criterion *Criterion 1 (C1): Distance of data to the classifier boundary and training data* gives the best results, 3) the global performance of informative/non-informative frames classifier is improving while enlarging training set, and 4) the LSH optimizes the computation process.

Table 4, shows that all three proposed schemes reduce the number of clicks. Even random order improves a lot with respect to the naive approach. This phenomenon can be explained by the fact that the frame color in a certain video are well correlated. From the results it can be concluded that the criterion 1 looks to be the best sorting criterion for interactive labeling. Intuitively, the samples that are far from the boundary are classified with high confidence. However, when dealing with the frames that are not similar to the ones in the training set (there are far from the one in the training set), the confidence gives uncertainty measure. Therefore, when introducing examples, where the classifier performs an error (user switches the label) to the model it is highly probable that the boundary will change using small amount of data.

Introducing new data into the training set improves the final classifier performance and reduces the error by 2% after 10 iterations of the algorithm (where each iteration is new video introduced in the training set) (Fig. 4). Furthermore, the LSH in average reduces the number of frames to check by more than 80%. This means that tested videos have about 20% of the frames that are “strange”. While inserting new frames into the classifier model, it can be seen, that at each

**Table 1.** Results

Video	#frames	#strange frames	#clicks		
			Criterion_1	Criterion_2	Criterion_3
Video1	35847	4687	103	103	147
Video2	51906	10145	211	213	316
Video3	52777	5771	270	270	376
Video4	56423	13022	86	90	151
Video5	55156	7599	68	68	131
Video6	33590	17160	381	389	617
Video7	17141	1072	8	8	39
Video8	26661	5437	88	97	151
Video9	14767	1006	28	28	76
Video10	22740	1993	63	63	110
Average clicks per video	-	-	1.5%	1.5%	2.9%

iteration some data that are not represented in the training set are being found. The conclusion that can be drawn is that in order to create a good training set for informative/non-informative frame classification, the number of 20 WCE videos is not enough.

## 5 Conclusions

In this paper a system that minimizes the user effort during the process of constructing a good representation of the WCE data is presented. The methodology is based on two steps: 1) the detection of frames that enrich the training set representation and thus should be labeled, and 2) the interactive labeling system that allows to reduce the user effort, in the labeling process, using an online classifier which sequentially learns and improves the model for the label proposals. The detection of frames that enlarge the data representation has been performed using LSH. The LSH method allows a fast processing for getting efficient results for data density estimation. Three different sorting polices are defined and evaluated for the online classification: 1) *Distance of data to the classier boundary and training data*, 2) *Data density*, and 3) *Random order*. It is shown that by using adapted sorting criteria for the data we can improve the label proposal process and in this way reduce the expert efforts. Finally, we have observed that enlarging the initial training set with the non represented frames from unlabeled videos we achieve an improvement of classification performance.

## References

1. Bashar, M., et al.: Automatic detection of informative frames from wireless capsule endoscopy images. *Medical Image Analysis* 14(3), 449–470 (2010)
2. Coimbra, M., Cunha, J.: MPEG-7 visual descriptors: Contributions for automated feature extraction in capsule endoscopy. *IEEE TCSVT* 16(5), 628–637 (2006)
3. Gionis, A., Indyk, P., Motwani, R.: Similarity search in high dimensions via hashing. In: *Proc. of the 25th ICVLDB, VLDB 1999*, pp. 518–529 (1999)
4. Igual, L., et al.: Automatic discrimination of duodenum in wireless capsule video endoscopy. In: *IFMBE Proceedings*, vol. 22, pp. 1536–1539 (2008)
5. Jung, Y., et al.: Active blood detection in a high resolution capsule endoscopy using color spectrum transformation. In: *ICBEI*, pp. 859–862 (2008)
6. Kang, J., Doraiswami, R.: Real-time image processing system for endoscopic applications. In: *IEEE CCECE 2003*, vol. 3, pp. 1469–1472 (2003)
7. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Gaussian processes for object categorization. *Int. J. Comput. Vision* 88, 169–188 (2010)
8. Malagelada, C., et al.: New insight into intestinal motor function via noninvasive endoluminal image analysis. *Gastroenterology* 135(4), 1155–1162 (2008)
9. Orabona, F., Keshet, J., Caputo, B.: The projectron: a bounded kernel-based perceptron. In: *Proc. of the 25th ICML 2008*, pp. 720–727 (2008)
10. Settles, B.: *Active learning literature survey*. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
11. Vilarino, F., et al.: Intestinal motility assessment with video capsule endoscopy: Automatic annotation of phasic intestinal contractions. *IEEE TMI* 29(2), 246–259 (2010)