

Progressive Learning for Interactive Surveillance Scenes Retrieval

Jérôme Meessen, Xavier Desurmont, Jean-François Delaigle
Multitel asbl
Mons, Belgium
jerome.meessen@multitel.be

Christophe De Vleeschouwer, Benoît Macq
Communications and Remote Sensing Lab. (TELE)
Université catholique de Louvain (UCL), Belgium
devlees@tele.ucl.ac.be, macq@tele.ucl.ac.be

Abstract

This paper tackles the challenge of interactively retrieving visual scenes within surveillance sequences acquired with fixed camera. Contrarily to today's solutions, we assume that no a-priori knowledge is available so that the system must progressively learn the target scenes thanks to interactive labelling of a few frames by the user.

The proposed method is based on very low-cost features extraction and integrates relevance feedback, multiple-instance SVM classification and active learning. Each of these 3 steps runs iteratively over the session, and takes advantage of the progressively increasing training set. Repeatable experiments on both simulated and real data demonstrate the efficiency of the approach and show how it allows reaching high retrieval performances.

1. Introduction

We address the problem of retrieving arbitrary scenes of interest within surveillance video sequences. The sequences are acquired with fixed camera, so that moving objects can be extracted based on a static background estimation and foreground extraction. Nowadays, such solutions for content-based retrieval of surveillance data are ever more required as the number of video surveillance systems and the amount of stored data drastically increase. The main reasons for searching video surveillance data are forensic, i.e. to look for a video evidence after an incident occurred, infrastructure maintenance, i.e. to extract statistics, and offline performance evaluation of automatic video

analysis systems.

Today's surveillance retrieval systems rely on offline a-priori knowledge about the events of interests. This knowledge is exploited to identify key-frames, as in [8] or [9], or to characterize the objects of interest with appropriate visual features, as in [7].

Here, the proposed system builds on low-level video features extractors without any assumption made about the target events nor about the number and configuration of moving objects of interest that users will request.

Moreover, the retrieval session is defined as an iterative classification process where the user reacts on the retrieval results, and the learning engine integrates the user feedback to improve the retrieval at the next step. The objective is to efficiently exploit the user feedback to increase the knowledge about the target concept. Such a progressive learning process must include two important components. First, the feedback of the user is appropriately integrated by the system at each iteration, which is known as relevance feedback [13, 11, 12, 10]. Second, the candidate scenes to be evaluated by the user are carefully selected so as to maximize the information gain at each step, and reduce the number of iterations of the training session. This is the process of active learning [13, 15, 14].

Relevance feedback has been widely studied for still images retrieval over the last years. Current techniques include visual features prioritisation and user query modification [10], though most of them use user's labels to build a training set for classification. In particular, support vector machines (SVM) have proven to be a powerful tool to classify images, integrate the user feedback, and enable simple but efficient active learning [16, 17].

Region-based image retrieval becomes much more complex when the user is asked to judge entire images (or video frames), while the learning is achieved on the image regions they consist of. In other terms, while features are available for the regions, the user expresses feedback on entire images, composed of sets of regions. This is referred to as a multiple-instance problem [18]. Most existing solutions assume the user is looking for only one single region. This does not match our application context because we are interested in recognizing events that are likely to be characterized by the occurrence of multiple elementary visual patterns[18, 19]. In [1], a SVM-based learning method, “MILES”, has been proposed to solve the multiple-instance problem without any assumption on the number of region to consider for the target class. Since we consider that several objects might be required to defined our target surveillance events, our method is based on this system, as reviewed in Section 3.1.

This paper presents a novel progressive learning scheme for efficient region-based retrieval surveillance scenes, integrating multiple-instance and active learning. At each step of the session, the new information obtained from the user is incorporated in the system to improve the classifier and guide future active learning steps. After introducing the retrieval workflow in Section 2, we detail the adaptive classification method in Section 3. The active learning process is presented in Section 4. Section 5 describes the conducted experiments and results, while Section 6 concludes this paper.

2. System Overview

We consider a set of video surveillance frames that have been pre-processed. The pre-processing includes the detection of moving objects as well as the extraction of low-level features describing each object. The event targeted by the user is characterized by the relative position of some of the extracted objects. The goal of the retrieval is to identify these particular frames within a minimum number of steps consisting in questioning the user about the relevance of well-chosen frames. A retrieval session then consists in an iterative process including querying the user, i.e. to ask him to label a few frames, inferring a classifier from the incremented training set, deducing the class of the unlabelled frames and, eventually, selecting new frames to be presented to the user at the next iteration. The classification is based on support vector machines (SVM) and will be detailed in Section 3. The retrieval workflow is depicted on Figure 1.

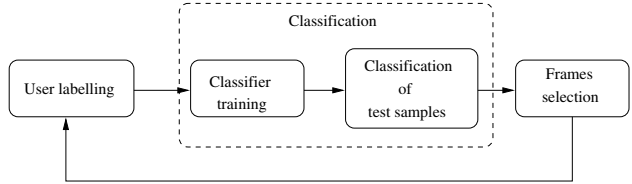


Figure 1. Flow chart of a retrieval session

3. Classification

Multiple-instance classification consists in labelling elements, called *bags*, based on the description of their sub-components, named *the instances*. The challenge comes from the ambiguous contribution of the instances that are not labelled directly and whose correspondence to the target concept is not known. The problem of classifying surveillance scenes using the description of their detected moving objects can thus be considered as a multiple-instance learning problem. While some of these objects may match the target class configuration, other may not.

Classification using multiple-instance learning has been already studied over the last years. In particular, methods have been presented based on support vector machines (SVM)[5]. However, most of them were restricted to target classes defined by a single instance. In [1], Chen et al. presented their *MILES* method to identify classes involving more than one instance. We review here this method and present in Section 3.2 how to extend MILES for interactive classification of surveillance scenes.

3.1. MILES overview

As defined in [1], MILES consists of two steps. First, by defining a similarity measure $s(B_i, x^k)$ between a bag B_i and an instance x^k , bags are mapped into a new space, where the multiple-instance problem becomes easier to express. Second, 1-norm SVM is used to calculate the hyperplane separating both the positive and the negative classes. These two steps are briefly reviewed in the following.

The similarity measure $s(B_i, x^k)$, is given by the shortest distance d between x^k and the instances x_{ij} of B_i :

$$s(B_i, x^k) = \max_j \exp \left(-\frac{d(x^k, x_{ij})}{\sigma^2} \right). \quad (1)$$

The exponential form of this similarity highly stresses the role of instances located in the neighborhood while other contributions are strongly reduced, the notion of neighborhood being regulated by σ .

This similarity enables the definition a new feature space \mathbf{F}_C where the multiple-instance problem becomes

more separable. Given a training set of l^+ positive bags and l^- negative bags, together presenting a total of n instances, a training bag B_i is mapped to a n -dimensional feature vector $\mathbf{m}(B_i)$ in \mathbf{F}_C using its similarity to each individual training instance. That is

$$\mathbf{m}(B_i) = [s(B_i, x^1), s(B_i, x^2), \dots, s(B_i, x^n)]. \quad (2)$$

\mathbf{m}_i^+ denotes the mapping coordinates of the i^{th} positive bag, and \mathbf{m}_i^- for the negative bags.

SVM classification in \mathbf{F}_C then consists in defining the hyperplane separating the positive and the negative bags so that the margin between the closest positive bag and the closest negative bag is as wide as possible. More precisely, the goal is to build a linear classifier

$$y = \text{sign}(\mathbf{w}^T \mathbf{m} + b) \quad (3)$$

where \mathbf{w} and b respectively define the coefficients and the bias of the hyperplane and \mathbf{m} the feature vector of a bag. In a non-separable case, and using 1-norm of SVM, the classifier (3) is computed based on the following lagrangian minimisation:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi, \eta} \quad & \lambda \sum_{k=1}^n |w_k| + \mu \sum_{i=1}^{l^+} \xi_i + (1 - \mu) \sum_{j=1}^{l^-} \eta_j \quad (4) \\ \text{s.t.} \quad & [\mathbf{w}^T \mathbf{m}_i^+ + b] + \xi_i \geq 1, i = 1, \dots, l^+, \quad (5) \\ & -[\mathbf{w}^T \mathbf{m}_j^- + b] + \eta_j \geq 1, j = 1, \dots, l^-, \quad (6) \\ & \xi_i, \eta_j \geq 0, i = 1, \dots, l^+, j = 1, \dots, l^-. \quad (7) \end{aligned}$$

Equ. (5) and Equ. (6) expresses the constraints related to correct classification of the training bags. ξ_i and η_j denote penalties allowed for the i^{th} positive and the j^{th} negative bags respectively. A penalty is positive when the distance between its corresponding bag and the classification plane is lower than 1.

μ balances the contributions between positive and negative bags which is particularly useful in case of strong disproportion between the size of their respective training sets. In our experiments, we have selected μ so that:

$$\mu \cdot l^+ = (1 - \mu)l^-. \quad (8)$$

In equation (4), λ trades off the amount of tolerated classification penalties and the risk of overfitting. λ and is often called the regularization parameter [2]. Nonetheless, classification penalties on the training bags, ξ_i and η_j , remain possible in a non-separable case.

The minimisation of Equ. (4) results in a set of non-zero w_k^* and an optimal value b^* that together define

the separating hyperplane and thus enable the classification of an unlabelled test bag:

$$y = \text{sign} \left(\sum_k w_k^* s(x^k, B_i) + b \right). \quad (9)$$

Intuitively, the non-zero w_k^* values of w , whether they are negative or positive, correspond to instances among the training bags, that are relevant for discriminating the positive class from the negative one in \mathbf{F}_C . In the following, we name these particular instances, the *reference instances*. The use of 1-norm of $|w_k|$ in equ. (4) is motivated by its sparsity properties. Given our assumption that the target concept involves only a few instances, the sparsity allows finding only a small set of representative instances.

3.2. Visual features selection and definition of the distance between instances

As described in Section 3.1, MILES enables to class bags as far as a similarity measure s and a set of parameters - μ and λ - are properly defined. In this section, we present how to estimate the similarity measure, i.e. the distance d between instances, in our application context.

3.2.1 Distance between two instances

In [1], a common Euclidean distance measure is used for estimating the similarity between two instances (Equ. (1)). In our extension of MILES, we rely on relevance feedback to weight each term of the Euclidean distance so that each feature can be independently favoured in order to boost the classification [6]. If V is the number of features used for describing an instance:

$$d(x^k, x_{ij}) = \sqrt{\sum_{v=1}^V u_v \cdot (x_{ij,v} - x_v^k)^2}. \quad (10)$$

The weights u_v are obtained by identifying the visual features that most efficiently discriminate between the positive and negative classes. A good approximation for u_v is provided by the ratio of the standard deviations of feature v among the negative and positive instances respectively.

3.2.2 Parameter σ

In [1], σ is tuned through cross-validation. We propose an alternative approach to estimate σ directly based on the distribution of the instances of training bags.

σ basically controls how similar two instances must be considered in the original feature space, see Equ.

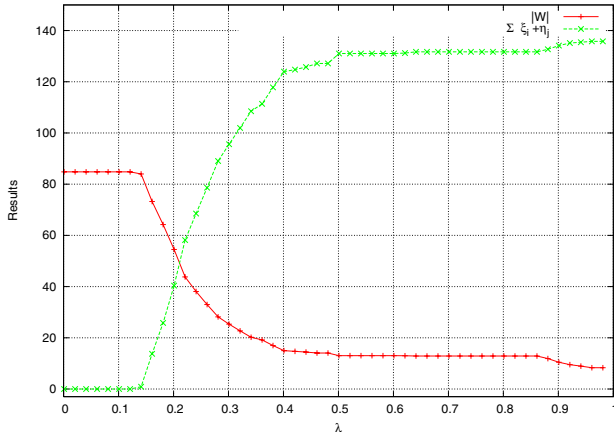


Figure 2. Example of regularization path, i.e. the evolution of the magnitude of the weights $|\mathbf{w}|$ and the penalties ξ, η as a function of λ

(1). In most practical cases, σ can be estimated in the original feature space by the allowed deviation around an average target instance. Moreover, given the reference instances used for the classification of bags, σ can also be approximated by the standard deviation of the positively weighted instances in equation (9) compared to the deviation of the other available instances.

3.3. λ -path study

In Equ. (4), λ is a critical parameter trading off the amount of classification penalties on the training samples and overfitting of the inferred classifier. As introduced in section 3.1, for a given classifier, the penalty ξ_i (or η_i) is non-zero for the training bags that are located within or at the wrong side of the SVM margin. Knowing from SVM theory that the margin's width is proportional to $1/|\mathbf{w}|$ [5], the larger λ , the smaller $|\mathbf{w}|$ and, consequently, the more bags stay within the margin, i.e. the larger $\sum_i \xi_i$ and $\sum_j \eta_j$. When λ is much too large, all bags are located inside an excessively wide margin ($|\mathbf{w}|$ becomes minimal). Contrarily, the lower λ , the narrower the margin and the lower the number of penalties from training instances. If λ is too small, the margin becomes too narrow and overfits the training data with the drawback of poor classification performances on any test set. In [2], the evolution of $|\mathbf{w}|$ and $\sum_i \xi_i + \sum_j \eta_j$ as a function of λ is referred to as the *regularization path*. An example is illustrated on Figure 2.

The definition of the λ value maximising the classification performances is not an easy task and is very dependent on the distribution of the data. Intuitively, if the classes overlap each others, there must be regularization, i.e. λ not too small.

An appropriate λ is traditionally selected within a

predefined grid of candidate values, based on cross-validation [1]. However this process can soon become heavy if no a-priori knowledge about the appropriate range of the relevant λ values is available. In our particular application scenario, each step of the iterative retrieval process include both labelling new training samples and the design of a new SVM classifier. Hence, the cost of a blind exhaustive grid search across a grid of potential λ should be avoided.

A heuristic value for the optimal λ has been proposed by Joachims in his implementation of SVM^{light} [3, 4]. But a single heuristic value obviously does not match all situations.

In [2], Hastie et.al. have provided an analytical study of the regularization path, i.e. the evolution of the model $[\mathbf{w}, b, \xi, \eta]$ as a function of λ . They showed that the regularization path is completely defined by a set of key λ values, named the *elbow points*. The study starts from an excessively high value of λ , where all instances are located inside the margin. When λ decreases, the norm of \mathbf{w} progressively increases and the margin consequently becomes narrower until a specific λ value where a first bag hits the margin frontier. This is the first elbow point. Each following elbow point basically corresponds to a new configuration of training bags with respect to the margin. Between two elbow points, the evolutions of both $\|\mathbf{w}\|$ and the sum of penalties are linear. Starting from the quadratic SVM formulation - i.e. using $\|\mathbf{w}\|$ rather than $|\mathbf{w}|$ in (4) - and its dual expression, they provided an analytical definition of each elbow point. The authors then suggested that parsing the regularization path can be achieved at relatively reduced cost based on these elbow points and the linear evolution between them.

Inspired by [2], we propose a method for guiding the crossvalidation procedure in our multiple-instance framework, derived from the 1-norm formulation. The approach allows efficiently scanning the regularization path in search for the λ values that should be checked through cross validation on the training set. The goal of this crossvalidation is then to identify the best λ value to be used for the next classification of the test samples.

When the regularization is excessive, all training samples are inside the margin and ξ_i and η_j are all strictly positive. In this situation, defining $C(\lambda, \mathbf{w}, b)$ as:

$$C(\lambda, \mathbf{w}, b) = \lambda \sum_{k=1}^n |w_k| + \mu \sum_{i=1}^{l^+} (1 - [w^T m_i^+ + b]) + (1 - \mu) \sum_{j=1}^{l^-} (1 + [w^T m_j^- + b]). \quad (11)$$

we may rewrite the minimisation of Equ. (4) as :

$$\min_{\mathbf{w}, b, \xi, \eta} C(\lambda, \mathbf{w}, b) \quad (12)$$

For a given λ , the partial derivative $\frac{\partial C(\lambda, \mathbf{w}, b)}{\partial |w_k|}$ is equal to:

$$\lambda - \text{sign}(w_k) \left[\mu \sum_{i=1}^{l^+} \left(\mathbf{m}_{i,k}^+ + \frac{\partial b}{\partial |w_k|} \right) - (1 - \mu) \sum_{j=1}^{l^-} \left(\mathbf{m}_{j,k}^- + \frac{\partial b}{\partial |w_k|} \right) \right] \quad (13)$$

Here, $\mathbf{m}_{i,k}^+ = s(B_i^+, x^k)$ and $\mathbf{m}_{j,k}^- = s(B_j^-, x^k)$.

As μ is chosen to compensate the disequilibrium between positive and negative samples,

$$-\mu \sum_{i=1}^{l^+} \frac{\partial b}{\partial |w_k|} + (1 - \mu) \sum_{j=1}^{l^-} \frac{\partial b}{\partial |w_k|} = 0, \quad (14)$$

and, defining S^k to be

$$S^k = \mu \sum_{i=1}^{l^+} m_{i,k}^+ - (1 - \mu) \sum_{j=1}^{l^-} m_{j,k}^-, \quad (15)$$

we can write:

$$\frac{\partial C(\lambda, \mathbf{w}, b)}{\partial |w_k|} = \lambda - \text{sign}(w_k) S^k. \quad (16)$$

Hence, as long as $\lambda \geq |S^k|, \forall k, \frac{\partial}{\partial |w_k|} \geq 0$ and there is no incentive for w_k to become $\neq 0$ since Equ. (4) must be minimized. Once $\lambda = \max_k |S^k|$, the first elbow point is hit ($\frac{\partial}{\partial |w_k|} = 0$) and w^i , with $i = \arg \max_k |S^k|$ becomes non-zero. Such non zero weights means that the corresponding instance x_i is used as reference in \mathbf{F}_C to separate the positive class from the negative one. Note that in this configuration, $B_i \ni x_i$ is the first bag to get out of the margin, and it no longer contributes to the penalty terms $\mu \sum_{i=1}^{l^+} \xi_i + (1 - \mu) \sum_{j=1}^{l^-} \eta_j$.

Consequently, the particular value

$$\lambda^* = \max_k |S^k| \quad (17)$$

is the first value of λ where a significant change appear in the regularization path. We thus consider this is the first candidate value for the crossvalidation test.

In order to identify other relevant λ , we make two assumptions. First, when a bag gets out of the margin, we consider it no longer returns inside it for a decreasing λ . Second, a bag containing instances that are very similar to reference instances from bags that are already out of the margin also gets out of the margin,

as far as they belong to the same class. Analytically, this means that their penalty terms are set to 0 in Equ. (12) and for all instances x_k that have not been selected yet, $\frac{\partial C(\lambda, \mathbf{w}, b)}{\partial |w_k|}$ equals:

$$\lambda - \text{sign}(w_k) \left[\mu \sum_{x_i \in R^+} m_{i,k}^+ - (1 - \mu) \sum_{x_j \in R^-} m_{j,k}^- \right], \quad (18)$$

where R^+ and R^- respectively denote the sets of positive and negative bags that are still considered to be located within the margin. This can still be written as:

$$\frac{\partial C(\lambda, \mathbf{w}, b)}{\partial |w_k|} = \lambda - \text{sign}(w_k) S^n, \quad (19)$$

with:

$$S^n = S^{n-1} - \text{sign}(w^k) \left[\sum_{i, B_i \notin \{R^+ \cup R^-\}} m_{i,k} + \sum_j m_{j,k} \right] \quad (20)$$

for all j such that B_j is very similar to the previously selected instances.

From S^n , we can then identify the next reference instance x_k with $k = \arg \max_k |S^n|$, providing the next elbow value for λ that should be explored during the crossvalidation. This process can be iteratively repeated to parse the whole regularization path. In case the training set is not consistent enough, these particular λ values may only describe the key angles of the regularization path. It is then easy to derive intermediate λ values to be investigated.

In our considered retrieval scheme, the definition of the λ values is achieved on the training set. However, it is worth noting that the size of the training set has an impact on the definition of these values. In particular, the classification results obtained with λ values computed based on the full training set may not correspond to those obtained with a subset of the training bags, as done during the search for the optimal λ through cross-validation. As illustrated in our experiments, this still an open issue that we are currently investigating.

As a conclusion, equations (17) and (20) provide a set of key λ values that must be explored during the cross-validation since they correspond to significant changes in the classifier inferred from each training set.

4. Active Learning

In the considered retrieval approach, the training set incrementally grows over the session thanks to the novel samples labelled by the user. Since classification performances always depends on the training set,

it is of prime importance to carefully select the samples to present to the user. In particular, the selected samples should maximize the information gain at the next classification step. This process is known as active learning.

In the case of SVM classification, Tong and Koller have demonstrated that the most informative samples were the closest to the separating hyperplane, which corresponds to the intuition since they are the most ambiguous samples [16, 17].

In our case, given the weight vector \mathbf{w}^* obtained from the last classification step, the most ambiguous bags B^M have instances with minimal distance to the separation plane:

$$B^M = \arg \min_i |\mathbf{w}^{*T} \mathbf{m}_i + b| \quad (21)$$

Knowing the number of bags to present to the user, this thus gives us a direct way to identify the bags to select.

5. Preliminary Experiments

5.1. Test data and features extraction

The proposed system has been tested on both simulated and real surveillance data. For the simulated data, each positive bag contains two instances. One of these instances presents two features falling in a gaussian distribution. The second instance is randomly placed in the feature space. Negative bags consist of two randomly located instances. Consequently, a bag is positive if it contains at least one instance from the gaussian distribution.

The real surveillance video sequences consists of a 20596 frames concatenation of the IEEE PETS 2006 video excerpts presenting realistic situations in a train station hall[20]. A mixture of gaussian method is used to extract the moving objects from the static background [21]. Note that no particular attention is paid to the illumination change between the original video excerpts, which leads to numerous redundant objects detection at the transition frames. For each moving objects, several visual features are computed: horizontal and vertical position in the frame, width and height, density - i.e. ratio of the actually ‘active pixels’ over the smallest rectangle area covering the object - and RGB colour histogram. The average number of detected objects (i.e. instances) per frame (i.e. bags) is about 4. A groundtruth has been manually established to identify several scenarios involving multiple objects, such as “activity at the restricted gate when somebody else is walking along the train”.

A retrieval session starts with a random selection of frames presented to the user who then labels a few of

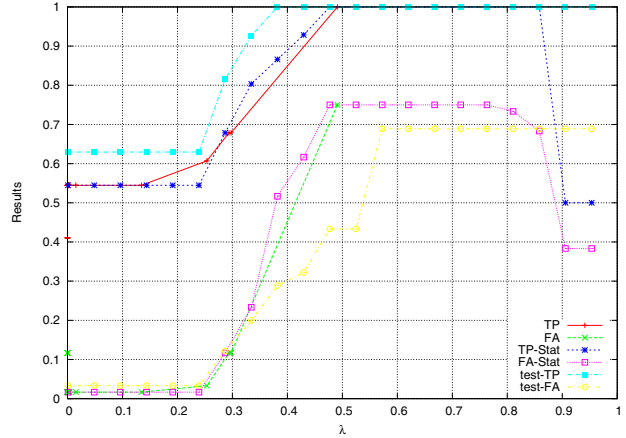


Figure 3. Comparison of performances selected after cross-validation over the training set (Stat) and on the test set (Test-Stat). TP and FA correspond to the λ values obtained with the proposed method on the simulated data.

them, either by simulation or actually. The number, f , of frames labelled at each step is fixed for the session and is a simulation parameter.

Three criteria are used to compare the method during test phases: the true positive rate (TP), i.e. number of correctly retrieved positive bags over the total number of target positive bags from the test set, the false alarm rate (number of false positive over total number of test bags) and the F_5 - score, i.e. an average of TP and precision (number of true positive over total number of retrieved bags)

$$(1 + \alpha) \frac{TP + Precision}{Precision + \alpha \cdot TP} \quad (22)$$

with α set to 5 since, in a surveillance scenario with rare target events, the true positive rate is far more important than the precision.

5.2. λ path study

Figure 3 compares the true positive rate and false alarm rate during a single step session with only one gaussian defining the positive bags. 150 bags have been simulated with 30% of positive bags, while 60 of them are used for the training. The experiment starts by analysing the training set. The elbow λ values are estimated with our method (see Section 3.3) and average detection (TP) and false alarm rate (FA) are computed via a 2-fold cross-validation within the training set. In order to compare these key values with a regular grid search, we repeated the cross-validation for a set of statically selected λ values. The corresponding cross-validation results are depicted by the ‘Stat-TP’ and ‘Stat-FA’ curves. It can be observed that the selected values are indeed representative of the regularization

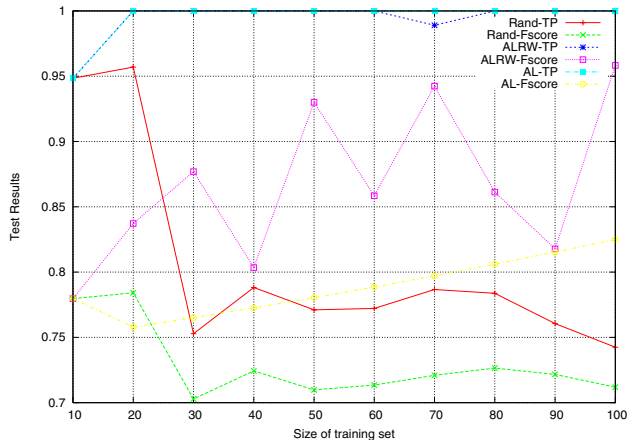


Figure 4. Comparison of true positive rate and F-Score with simulated data: randomly selected bags (RAND), Active Learning (AL) and Active Learning together with reweighting of visual features (ALRW)

path. This confirms that our method enables a faster scanning of the candidate λ values.

In a second phase, we compare the cross-validation results with those obtained on the unlabelled test set. Here, we again conducted the test for a regularly increasing λ . The results are represented by 'Test-TP' and 'Test-FA'. These curves indicate that the test results roughly match those computed from the cross-validation on the training set. However, as mentioned in Section 3.3, for a given λ value, the classification of the test set is achieved with the full training set, which is twice as large as the sub-training sets used during the cross-validation. This leads to a difference of classification results between the training and the test sets.

5.3. Results for a retrieval session

Figure 4 presents the test results for a 10 steps retrieval session on the simulated data. 30% of the 300 original bags are positive and the λ value is fixed to 0.4. We compare 3 methods for selecting frames to be labelled by the user: random selection (RAND), active learning (AL) and active learning combined with reweighting of the visual features (ALRW). Both the detection rate (TP) and F-score are presented. As illustrated, both the AL and ALRW method quickly succeed in reaching 100% of detection, which is not the case in the case of randomly selected frames. The F-score curves show that the active learning provides better results than the random selection but significant improvement is only obtained when visual reweighting of the features is added. It can also be noticed that despite the good ALRW detection performances, the corresponding F-score strongly oscillates. An efficient way to smooth it is part of our future work.

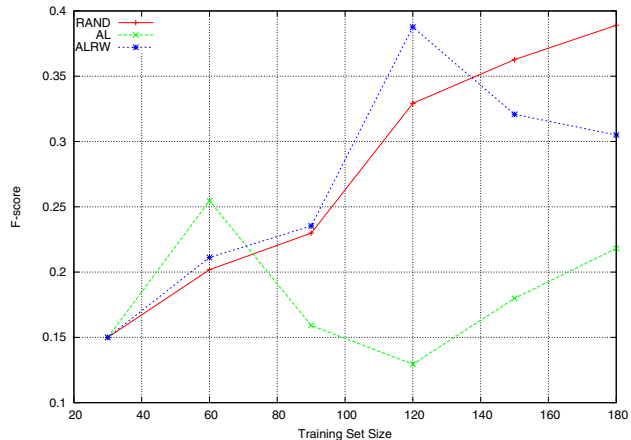


Figure 5. Comparison of F-score from 3 frames selection methods on real IEEE PETS sequence

Figure 5 presents the F-score results obtained on the real data. The 3 selection methods are compared. We see that the active learning leads to poor results compared to the 2 others. Our explanation is that, compared with a random selection of bags, this method doesn't explore the \mathbf{F}_C space enough to actually localize the best reference instances. We consider this as a way to improve the method. Nonetheless, when combined with features re-weighting, ALRW strongly enhances the results, even though oscillation is still to be observed.

6. Conclusion

An interactive retrieval system adapted to the particular constraints of surveillance video data has been presented. It integrates visual features reweighting, multiple-instance 1-Norm SVM classification and active learning. Each of these steps takes into account the dynamically increasing training set.

In particular, we have shown how the search for an optimal regularization parameter λ can be guided at each retrieval step based on the training data distribution. Moreover, relevance feedback is integrated within a multiple-instance framework.

Preliminary experiments have demonstrated the efficiency of the approach on both simulated and publicly available real data. 3 future improvements of the system have been identified. First, to automatically adapt the regularization parameter to the training set size. Second, to smooth the evolution of the active learning when combined with feature re-weighting. Third, improving the active learning by exploring the instances space at the firsts retrieval iterations.

References

- [1] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-instance Learning via Embedded Instance Selection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no.12, pp. 1931-1947, December 2006
- [2] T. Hastie, S. Rosset, R. Tibshirani and J. Zhu, "The Entire Regularization Path for the Support Vector Machine, ", *Journal of Machine Learning Research* (0), pp. 1-24, 2004.
- [3] T. Joachims, "Making Large-Scale SVM Learning Practical, " in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press, 1999.
- [4] T. Joachims, "Estimating the Generalization performance of an SVM Efficiently, " *proc. of International Conference on Machine Learning (ICML)*, 2000.
- [5] V. Vapnik, "The Nature of Statistical Learning, " Springer-Verlag, 1996.
- [6] S. Askoy, R. M. Haralick, F. A. Cheikh and M. Gabbouj, "A Weighted Distance Approach to Relevance Feedback," *Proc. of 15th International Conference on Pattern Recognition*, vol. 4, pp. 812-815, September 2000.
- [7] C. Kim and J.-N. Hwang : Object-Based Video Abstraction for Video Surveillance Systems. *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 12, No. 12, pp. 1128-1138, December 2002.
- [8] E. Stringa, C.S. Rgazzoni: "Real-Time Video-Shot Detection for Scene Surveillance Applications," *IEEE Trans. on Image Processing*, Vol. 9, No.1, pp. 69-79, January 2000.
- [9] A. Hampapur, L. Brown, J. Connell, M. Lu, H. Merkl, S. Pankanti, A.W. Senior, C. Shu, and Y-L Tian: "The IBM Smart Surveillance System," in *IEEE Conference on Computer Vision and Pattern Recognition*, Washington D.C., June 2004
- [10] X.S. Zhou and T.S. Huang: "Relevance Feedback in image Retrieval: A Comprehensive Review," *Multimedia Systems*, vol. 8, pp. 536-544, Springer-Verlag, 2003
- [11] Y. Rui, T.S. Huang, M. Ortega and S. Mehrotra: "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval," *IEEE trans. on Circuits and Systems for Video Technology*, vol. 8 (5), pp; 644-655, September 1998
- [12] X. He, O. King, W.-Y. Ma, M. Li and H.-J. Zhang: "Learning a Semantic Space from User's Relevance Feedback for Image Retrieval," *IEEE. tran. on Circuits and Systems for Video Technology*, vol. 13(1), pp. 39-48, January 2003.
- [13] I.J. Cox, M. L. Miller, T. P. Minka, T.V. Papathomas and P.N.Yianilos: "The Bayesian Image Retrieval System, PicHunter: Theory, Implementation, and Psychophysical Experiments," *IEEE trans. on Image Processing*, vol 9(1), January 2000.
- [14] Y. Freund, H.S. Seung, E. Shamir and N. Tishby: "Selective Sampling Using the Query by Committee Algorithm," *Machine Learning*, 28, pp. 133-168, 1997.
- [15] C. Zhang and T. Chen: "An Active Learning Framework for Content-Based Information Retrieval," *IEEE trans. on Multimedia*, vol 4, pp. 260-268, 2002.
- [16] S. Tong and D. Koller: "Support Vector Machine Active Learning with Application to Text Classification," *Journal of Machine Learning Research*, 2, pp. 45-66, 2001.
- [17] S. Tong and E. Chang: "Support Vector Machine Active Learning for Image Retrieval," *proc. of 19th ACM International Conference on Multimedia*, pp. 107-118, Ottawa, 2001.
- [18] T.G. Dietterinch, R.H. Lathrop and T. Lozano-Perez: "Solving teh Multiple Instance Problem with Axis-Parallel Rectangles," *Artificial Intelligence*, vol. 89(1-2), pp. 31-71, 1997.
- [19] C. Zhang, S.-C. Chen and M.-L. Shyu: "Multiple-Object Retrieval for Image Database Using Multiple Instance and Relevance Feedback," *proc. of IEEE International Conference on Multimedia and Expo (ICME)*, vol 2, pp. 775-778, Tapei, Taiwan, June 2004.
- [20] *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, New-York, USA, June 2006.
- [21] Stauffer C., Grimson W.E.L.: "Adaptive background mixture models for real-time tracking," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 246-252, June 1999.