User-Centric Retrieval of Visual Surveillance Content

Jerome Meessen, Xavier Desurmont, Christophe De Vleeschouwer and Jean-François Delaigle

Abstract—**An interactive retrieval method adapted to surveillance video is presented. The approach is formulated as an iterative SVM classification and builds upon the two major specificities of the surveillance context, namely the multiple instance nature of the data and the reduced number of training examples the user can provide at each round. The later issue is solved thanks to a new adaptive active learning strategy as well as an intuitive graphical user interface. The system has been validated on both synthetic and real datasets.**

Index Terms—**Surveillance video retrieval, multiple-instance, relevance feedback, active learning, GUI.**

I. Introduction

THIS paper addresses the problem of retrieving arbitrary scenes of interest within surveillance video sequences stored in a database. A scene is described by one or several video frames with a common configuration of (some of) their foreground objects. The goal of our system is to learn which particular configuration of objects the user is looking for, while minimizing his/her work load.

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. Since there is no a priori knincreasing training set. In other words, compared to existing surveillance video retrieval systems, we adopt a versatile content-based approach, guided by the interactions with the user, rather than relying on the indexation of pre-defined scenarios.

Our retrieval method is modelled as an iterative supervised classification problem. A session is composed of a few steps repeated until the retrieval performances satisfy the user. This is depicted on Figure 1.

First, the user labels a few frames. This operation is kept very simple: the user only tells the system whether the presented frames are matching his/her target scene or not. The number of frames labelled at each round is expected to be small, i.e. typically 5 to 10.

The labelled frames are used to train a soft-margin SVM (Support Vector Machine) classifier, based on features extracted beforehand, either at low/mid visual level or at higher semantic level. The classifier is then exploited to predict a label for the unknown frames as well as their similarity to the inferred target class.

Finally, the system carefully selects the next frames that should be presented to the user for labelling.

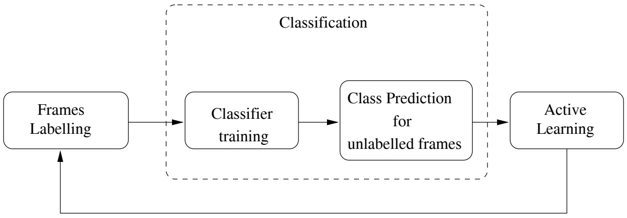

Fig. 1. Flow chart of a retrieval session

The specific constraints of the surveillance context led us to face two major problems:

1) The nature of the surveillance data imposes a specific classification approach. Here, we present a multiple-instance framework, which allows classification of the data.
2) The load on the user must be minimized. So, the system must learn as fast as possible what the target class is, while the number of training examples is small and slowly growing. This is achieved by carefully selecting the frames to be presented to the user. In this paper, we propose an online adaptation of the active learning strategy, taking advantage of the progressing training set. Moreover, a graphical user interface is designed so as to provide intuitive navigation through the retrieval results. It allows the user to easily label groups of frames and consequently boost the classification performances.

II. Multiple-instance framework

Surveillance video data have a very specific nature compared to classical images. Basically, through background learning, foreground objects can be extracted and described by features such as position, texture, shape, direction etc. However, the number of objects per frame can be very large in noisy conditions and varies a lot from frame to frame. Moreover, not all objects truly correspond to real-life objects, vehicles or individuals.

The classification of surveillance video frames then is formalized as a multiple-instance problem, were the frames are the *Bags* -, while the foreground objects are the - *instances*[3], [2]. Since the number of instances per bag is highly variable, there exist an ambiguity about their relevance to the received label. While a few instances may justify a positive label, other instances from the same bag may indeed be totally irrelevant. In [1], we presented a framework to represent the data and enable a SVM classification.

It consists of two steps. First, we define a similarity measure $s(B_i, x^k)$ between a bag B_i and an instance x^k, which is used

J. Meessen, X. Desurmont and J.-F. Delaigle are with Multitel, Beligum. jerome.meessen@multitel.be

C. De Vleeschouwer is with the TELE lab. of UCL, Belgium

to map the bags into a new space, where the multiple-instance problem can be solved. Second, in this space, a soft-margin SVM is trained to predict the label of the unlabelled bags:

$$y = \text{sign}\left(\sum_k w_k s(x^k, B_i) + b\right), \tag{1}$$

were k is the number of instances among the training bags. w_k and b are the decision plane parameters provided by the SVM training.

III. Adaptive Active Learning

Successfully achieving a fast learning when only a few training examples are provided at a time is only possible by carefully selecting the frames to be presented to the user. The goal is to maximize the information gain these new examples will offer, if the user labels them. Such a selection process is called active learning and is a well known challenge in information retrieval[4], [5].

Three traditional approaches are commonly used for selecting new examples. First, the most classical technique consists in selecting, the most uncertain examples. It has been shown that the proximity to the decision plane is a good approximation of such uncertainty. Receiving labels for these examples indeed allows refining the classification hyperplane.

In our case, given the weight vector $\mathbf{w}$ obtained from the last classification step (Equ.(1)), the most ambiguous bags B^u, with minimal distance to the separation plane, are defined by:

$$B^u|u = \arg\min_i |\mathbf{w}^T\mathbf{m_i} + b| \tag{2}$$

The second technique intends to gather knowledge about the data distribution in the feature space. In other words, its goal is to explore the space in order to discover all "clusters" of instances. Intuitively, this should be achieved before trying to refine the decision plane. The dynamic combination of such exploration and the refinement of the hyperplane is discussed in [6], [7].

Nonetheless, the exploration starts being useful only when a minimum of positive examples are known so that a first rough decision plane can be estimated. This is the goal of the third active learning approach. In our specific case, given the possible disproportion between the classes and the small number of training exampes provided at each round, it is not obvious that random exploration will be efficient at the start of the retrieval session.

Here, our proposed approach dynamically combines these three active learning methods. It first starts by looking for new positive examples by selecting the most certain examples. The active learning then explores the feature space and finally refines the decision plane using Equ.(2).

IV. Graphical user interface

The retrieval performances, as every classification task, is very dependent on the training set. Even though active learning allows dealing with small training sets, we developed an intuitive graphical user interface (GUI) that helps the user to quickly and easily label video frames. The GUI includes several tools and functionalities such as navigation through and multiple resolution display of the retrieval results. A speparate window invites the user to check the frames selected by the active learning engine (see Figure 2).

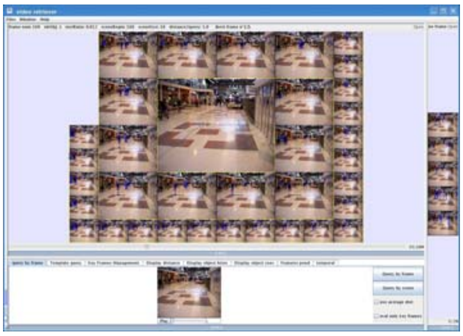

Fig. 2. Snapshot of the retrieval interface. The right hand window presents examples selected by active learning.

V. Conclusion and future work

A new approach for interactive retrieval of CCTV content is presented. It has been validated and compared to state-of-the art methods on both synthetic multiple-instance data and realistic reference data from the IEEE PETS workshop[10]. Future work includes the use of new modalities like voice or gesture recognition to help the user interacting with the results. This is in line with our initial philosophy stating that the user's task must be intuitive and simple.

Acknowledgment

This work is partially funded by the Walloon Region project IRMA [8] and by the European FP6-CARETAKER[9]. The experiments dataset includes the IEEE PETS 2006 video excerpts[10].

References

[1] J. Meessen, X. Desurmont, J.-F.Delaigle, C. De Vleeschouwer and B. Macq, "Progressive Learning for Interactive Surveillance Scenes Retrieval, " *in Proc. of IEEE Computer Society conference on CVPR 2007, 7th IEEE International Workshop on Visual Surveillance (VS 2007)*, Minneapolis, USA, 22 June 2007

[2] 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

[3] T.G. Dietterich, R.H. Lathrop and T. Lozano-Perez: "Solving the Multiple Instance Problem with Axis-Parallel Rectangles," *Artificial Intelligence*, vol. 89(1-2), pp. 31-71, 1997.

[4] Y. Freund, H.S. Seung, E. Shamir and N. Tishby: "Selective Sampling Using the Query by Commitee Algorithm," *Machine Learning*, 28, pp. 133-168, 1997.

[5] 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.

[6] Y. Baram, R. El-Yaniv and K. Luz: "Online Choice of Active Learning Algorithms," *Journal of Machine Learning Research*, 5, pp. 255-291, 2004.

[7] T. Osugi, D. Kun and S. Scott: "Balancing Exploration and Exploitation: a New Algorithm for Active Machine Learning," *proc. of Fifth IEEE International Conference on Data Mining (ICDM'05)*, pp. 330-337, 2005.

[8] Walloon Region Project "IRMA - Multi-modal Interfaces for Retrieval of audio-visual Archives," project website: http://www.irmaproject.net/

[9] EU FP6 project CARETAKER. Website: http://www.ist-caretaker.org

[10] IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, New-York, USA, June 2006.