Fast Support Vector Classifier for Automated Content-based Search in Video Surveillance

Cătălin A. Mitrea[1], Ionuț Mironică[1], Bogdan Ionescu[1,2], Radu Dogaru[1]
[1] LAPI & Natural Computing Labs, University “Politehnica” of Bucharest, 061971, Romania
[2] LISTIC, University Savoie Mont Blanc, 74940 Annecy-le-Vieux, France
Email: *catalin.mitrea@uti.ro*, {*imironica,bionescu*}*@imag.pub.ro*, *radu_d@ieee.org*

***Abstract*- In this article we present and test a specialized classifier, i.e., Fast Support Vector Classifier (FSVC), which is employed for multiple-instance human retrieval in video surveillance. Thanks to its low complexity and high performance in terms of computation and speed, FSVC is adapted to ease the generalization of the feature space using only a limited number of samples in the training process. To validate the performance, FSVC is evaluated on two standard video surveillance datasets. It obtains superior or similar results in terms of precision and recall compared to the close related state-of-the-art Support Vector Machines approaches.**

I. Introduction

The importance of automated video surveillance is almost tangible – it increases efficiency of human operators while reducing the number of possible threats to continuously increasing urban population. Recent world events have prompted major players to redesign the concepts of physical security. Due to relatively ease of access to low cost hardware and high processing power, intelligent video surveillance gained large attention during last years and found rapidly applications in industry fields such as public safety. On these premises, a high performing automated video surveillance becomes essential.

As a basis for intelligent video surveillance, effective multiple-instance based object (e.g., humans) retrieval is a difficult and complex task due to the large volumes of stored footage and the variable conditions in which the target objects are recorded, e.g., recorded from different perspectives (multiple sources), different weather conditions, different setups (e.g., indoor vs. outdoor) and appearances, noise, etc. Although all mentioned conditions play an essential part in system's performance, the selection of the processing approach to solve the task is also critical due to current technological limitations.

One limitation of the current state-of-the-art classifiers is low generalization power when only few training samples are available. This is a particular issue in video surveillance, as most of the time, the available instances of the target object to be searched for is limited, i.e., only few seconds of footage, or only few images are available to formulate the query.

Another issue is the processing time of the classifiers which should cope with real-time scenarios, running time which is often in relation with algorithm processing complexity. The speed is usually playing an important role in an efficient retrieval. As the amount of data to be evaluated is usually huge, the performance of the classifier in terms of processing time could be an impediment, thus making it unsuitable for real-time applications.

With respect to the aforementioned limitations and drawbacks, the main goal of this article is to introduce an alternative to the state-of-the-art SVM classifiers, previously referred to as RBF-M [14]. Its adaptation to this new video surveillance scenario is herein nominated as Fast Support Vector Classifier (FSVC). It is specialized for content-based search of humans in video surveillance datasets.

The reminder of the paper is organized as follows: In Section II several relevant automated video surveillance approaches are identified in the context of our current work. The proposed content-based search system is presented in Section III while its performance as results from various experiments is investigated in Section IV. Finally, Section V provides a brief summary and concludes the paper.

II. Related Work

A significant number of methods are reported in literature relating to effective automated video surveillance methods and their impact on physical security [1,2]. Most of the contributions are reported in the framework of finding automatic ways of describing video content while maximizing their representation power. They focus on understanding video contents using visual and spatio-temporal information [3]. Generally, all approaches exploit efficient content description and retrieval schemes. Many image feature extraction algorithms have been proposed and evaluated, e.g., shape [4], texture [5], color [6] or the popular feature point descriptors [7]. Their success is empowered by high invariance to rotation, change of scale, perspectives, illumination shifting or even signal perturbations.

Other approaches are investigated in the framework of decision making. At this level, the decisions are usually powered by classifiers. Some of them are investigating methods of automatic pre-processing and refinement of input data in order to leverage classifier precision [8]. For instance, in [9] the authors model the relationship between low-level events (concepts) in a framework based on latent SVM. Anyway, most of the efforts of classifier optimization are channeled to parameter optimization during the training process [10,11,12]. Almost all of the described methods are focusing on current available classifiers in literature while very few of them are proposing new or adapted approaches.

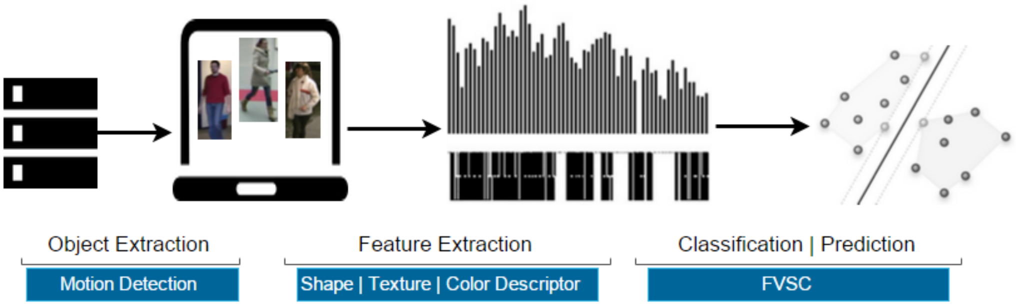
Object Extraction
Motion Detection
Feature Extraction
Shape | Texture | Color Descriptor
Classification | Prediction
FVSC

Figure 1. Proposed system architecture.

The current research focuses mainly on addressing the performance of the classification process by adopting low complexity and fast predictors. Current paper contributes further to the work in [13] by implementing another classifier and conducting further experiments on new dataset and validation scenarios. Findings and output of this research are contributed to steady advances to specialized classifiers for automated content-based search in large video sets acquired from video surveillance.

III. Proposed System

The proposed system is composed of three different layers: object extraction layer - based on motion detection, feature extraction layer - based on video description, and prediction layer - based on the FSVC reasoning (see Figure 1). In the first step, the system extracts the objects from the video frames using motion detection (a background subtraction algorithm based on Gaussian mixture models - the motion of each track is estimated by a Kalman filter). Secondly, a set of state-of-the-art feature extractors are employed for shape, texture and color information description. Finally, the proposed FSVC classifier is employed to label the input data and return content-based retrieval results to the end-user.

A. *The FVSV classifier*

The Fast Support Vector Classifier (FSVC) was first introduced in [14] (RBF-M - Modified Radial Basis Function Network). The architecture is based on simple arithmetic operators and employs simple Least Mean Squares (LMS) training in an expanded feature space generated by Radial Basis Function (RBF) kernels centered on support vectors selected via a simple algorithm. Unlike in the SVM, where sophisticated mechanisms are needed to identify support vectors, the FSVC needs only one single epoch to select the support vectors among feature vectors in the training samples. Then, simple Adaline training is performed in the expanded space formed of RBF kernels centered on the previously discovered support vectors. The FSVC model may be regarded as a RBF network with the output calculated as [15]:

$$y = w_0 + \sum_{k=1}^{m} w_k K(x, x^{j_k}) \tag{1}$$

where $w_0, \dots, w_m$ represents the weights of an outputted *Adaline* trained with LMS and x^{j_k} represents the center vector selected as the sample j_k from the *N* training samples. A sample of either test or training set is defined as a pair (x^k, d^k), where $x^k$is an input vector (scaled between [0; 1]) with size *n* and $d^k$is the training label that belongs to the set {-1, 1}, $d^k = 1$ indicating that x^k belongs to the search class.

In any given classification scenario, both training (TR) and testing (TS) sets are available. The TR is used in one single epoch to select a set of *m* support vectors using the following support vector selection algorithm:

```
m ←1; k ←1; $j_k$ ←1; // select the center of the first RBF unit as the first sample of the dataset;
ov ← 1; // set overlapping coefficient between two RBF units;
for j=2 to N // all training samples
        act ← $\sum_{k=1}^{m} K(x^j, x^{j_k})$ ; // compute activity level
        if (act < ov)
        k ← k +1; $j_k$ ← j; m ← m+1; // create a new RBF unit
        end if
end for
```

The result of the above algorithm is a set of indexes $j_k$indicating which of the training samples are selected as RBF centers. Training of the FSVC involves basically the following steps:

- Determine the RBF units using the above algorithm;
- Initialize with 0 all weights values of the Adaline (eq.1) and initialize the percentage of incorrect classification – PIC with a large value (e.g., 50);
- For the predefined *Ne* number of epochs alternate (eq.1) Adaline offline training with testing (i.e., computing the PIC for the test set). If PIC < reference (best) PIC value, refresh both PIC and best weights accordingly.

B. *Computational complexity*

A detailed analysis of FSVC (RBF-M) versus SVM was performed in [21] for several benchmark classification problems (single and multiple-class). Compared to SVM, the computational complexity is reduced for the following reasons:

- The training is much simpler, i.e., Adaline training, while only one epoch suffice to identify the support vectors; In practice several tens of epochs suffice to reach the maximal performance;
- Unlike in the SVM, where kernels must satisfy the Mercer's condition there is no such restriction for the FSVC. Consequently, simple triangular kernels may replace Gaussian ones and Manhattan distances may replace the Euclidian one with no significant performance loss;
- Unlike SVM where for multiclass problems different sets of support vectors are generated for each class, in the FSVC there is only one set of support vectors (and the same number of RBF-

units) for all classes, since each class is only assigned a different output Adaline (with the same nonlinear kernel for all classes). This results in a significantly lower number of kernel units than in the case of SVM. The effect is a much more compact classifier structure;

- All the above makes FSVC a very attractive architecture for real-time task implementations in automated video surveillance. This solution has the advantage of a significantly lower cost of implementation due to simple arithmetic modules which map conveniently in common digital but also analogue implementation technologies.

IV. Experimental Results

To test the proposed retrieval framework, experiments were conducted on two standard video datasets: SCOUTER[1] and PEVID-HD[2]. SCOUTER contains 30 videos and approximately 36,000 manually labeled frames (~10 fps, 704 x 675 pixels) and PEVID-HD is composed of 21 video clips and approximately 17,000 manually annotated frames (~25fps, 1920 x 1080 pixels). Selected datasets are rising particular video surveillance challenges due to the diversity of footages - shifting perspectives from camera to another (multiple source CCTV cameras), different weather conditions, different setups/lighting conditions and large variations of subject to be found (a total of 16 scenarios involving people are labeled).

A. System evaluation

Given the current task, i.e., content-based search in video surveillance, we consider *Recall* (accounts for the non-detections, FN) weight higher than *Precision* (accounts for the number of false positives, FP), i.e., we are interested in retrieving all of the existing instances while the number of false positives is of lower importance. We assess performance with the F_2 - score is computed as:

$$F_2 = 5\frac{PrecisionRecall}{4Precision+Recall} \qquad (2)$$

where *Precision* is computed as TP/(TP+FP) and *Recall* is TP/(TP+FN), where TP are the true positives (good detections). In practice a high rate of *Precision* denotes that the FSVC[3] classifier returned substantially more relevant results than irrelevant, while a high value of *Recall* concludes that the FSVC classifier returned most of the relevant results.

B. Feature extraction

Competitive results have been obtained using these four descriptors on other surveillance datasets, selected also for providing complementary discriminative power:

- *Local Binary Pattern (LBP - 256 dimensions)* [17] has become an attractive approach in various applications, mainly because of its discriminative power and computational simplicity. LBP is powered by a simple texture operator which labels the pixels of an image by thresholding the neighborhood of each pixel using a pattern and generating the result as a binary number;
- *Histogram of Oriented Gradients (HoG - 81 values)* - [18] very popular for shape based representations, those features exploits local object appearance within an image via the distribution of edge orientations. The image is divided into small connected regions (cells) and for each of them building a pixel-wise histogram of edge orientations is computed. In the end, the combination of these histograms represent the final descriptor;
- *Color Naming histogram (CN - 11 dimensions)* describes the global color contents and uses the Color Naming (CN) Histogram proposed in [19]. We select this feature, instead of the classic color histogram, because the color naming histogram is designed as a perceptually based color naming metric that is more discriminative and compact. It maps colors to 11 universal color names: ”black”, ”blue”, ”brown”, ”grey”, ”green”, ”orange”, ”pink”, ”purple”, ”red”, ”white” and ”yellow”;
- *Color moments (CM - 225 dimensions)* [20] provide a method for effective measurement for color similarity between image samples. There are three central moments of an image’s color distribution: mean, standard deviation and *skewness*. The image is divided in a 5x5 grid, and a color moment descriptor is computed for each individual cell.

C. Parameter tuning

In order to achieve the best generalization performance for proposed FSVC, different training parameters need to be properly adjusted. First of all, we apply a Principal Component Analysis (PCA) transform on the feature space. In Figure 2 is represented the accuracy of the proposed classifier during the training process. For the given feature input distribution best results are obtain for a PCA dimension ranging between 15 and 30. Regarding the FSVC kernel, one of the important parameters influencing the generalization performance is the radius *r,* which must be heuristically optimized, much like the “gamma” parameter in the Gaussian kernel of SVM. This optimization was done by exhaustive searching (see Figure 2) within a large interval [0.5;5]. Usually a rough search is first performed to locate a region where a finer search is performed next.

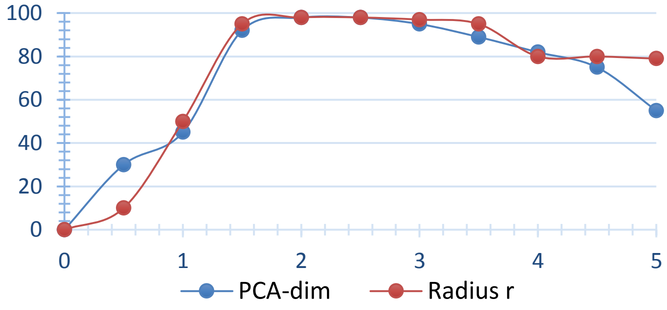

Figure 2. Variation of FSVC precision (%) during training in relation to the PCA dimension (x10) and Radius *r* (spreading radius of the activation function).

D. Comparision with SVM

Experimental results obtained after evaluation of FSVC and SVM are described in Table 1.

[1] http://uti.eu.com/pncd-scouter/rezultate.html

[2] http://mmspg.epfl.ch/pevid-hd

[3] http://www.mathworks.com/matlabcentral/fileexchange/49695-fast-support-vector-classifier--a-low-complexity-alternative-to-svm-

Table 1. F2 score evaluation for FSVC and SVM (best results with bold)

	F2-Score (%)			
Clasifier	SVM		FSVC	
Database	SCOUTER	PEVID-HD	SCOUTER	PEVID-HD
HoG	41.03	40.19	**42.48**	**47.37**
CM	38.87	45.00	**41.77**	**46.52**
LBP	42.26	47.26	**44.70**	**52.74**
CN	37.64	34.53	**38.35**	**41.10**
Fused	**44.08**	**48.56**	40.20	34.10

In SVM's case, experimental results are obtained with a non-linear RBF kernel and was selected as is denotes top performance in the literature on different image classification tasks. It can be observed that best F2-score is obtained by FSVC – LBP pair (52.74% on PEVID-HD dataset) and (44.7% on SCOUTER dataset). This is to the fact that RBF descriptor is a powerful feature for texture classification. While combined with proposed FSVC classifier, an efficiently nonlinear classification is performed which is suitable for our current task. Lower performance is obtained by HoG-FSVC pair while lowest scores are obtained by color descriptors (CN3x3 – SVM pair - 34.53% on PEVID-HD dataset). One reason is the diversity of the scenarios (i.e. different clothes colors of the people to be searched). Although FSVC performs well with individual descriptors, it denotes less performance than its SVM counterpart when feature space is fused. This is happening as SVM is less sensitive to dimension and noises of data input space.

V. CONCLUSION

The FSVC classifier is considered as a low complexity alternative to SVM for the use of multiple instance human retrieval task. Evaluation results revealed similar or better performance when compared with state-of-the-art classifiers as Support Vector Machines. While it denotes low computational complexity (kernel powered by summation and absolute value operations) it obtains high performance on decisioner assignment and classification of information extracted from video surveillance datasets. Although the FSVC obtains better results that SVM on both selected databases, it shows a sensitivity of the performance to the range of the analyzed input data. Its performance is decreasing considerably without a PCA transform on the input data.

Future work will address and investigate techniques to enhance further FVSC performance by employing specialized methods as co-training which are adapted to the situation when very few training samples are available.

ACKNOWLEDGMENT

This work has been supported under ExcelDOC POSDRU/159/1.5/S/132397, InnoRESEARCH POSDRU/159/1.5/S/-132395, PN-II-PT-PCCA-290/2014 and PN-II-IN-DPST-28DPST/30.08.2013.

REFERENCES

[1] A.J. Lipton, C.H. Heartwell, N.Haering, D. Madden, "Critical Asset Protection, Perimeter Monitoring, and Threat Detection Using Automated Video Surveillance," *In 36th Annual International Carnahan Conference on Security Technology*, 2002.

[2] N. Baaziz, N. Lolo, O. Padilla, F. Petngang, "Security and privacy protection for automated video surveillance,", *IEEE International Symposium on Signal Processing and Information Technology*, pp:17 - 22, 2007.

[3] B. Benfold, I. Reid, "Stable multi-target tracking in real-time surveillance video," *In CVPR*, pp. 3457-3464, 2011.

[4] Y. Mingqiang, K. Kidiyo, R. Joseph, "A Survey of Shape Feature Extraction Techniques", *in Pattern Recognition*, pp. 43-90, 2008.

[5] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "ContentBased Image Retrieval at the End of the Early Years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, nr. 12, pp 1349-1380, 2000.

[6] B. Ionescu, C. Rasche, C. Vertan, P. Lambert, "A Contour-Color-Action Approach to Automatic Classification of Several Common Video Genres", *in Springer-Verlag LNCS - Lecture Notes in Computer Science*, Eds. M. Detyniecki, P. Knees, A. Nurnberger, M. Schedl and S. Stober, vol. 6817, pp. 74-88, 2011.

[7] D. G. Lowe, "Distinctive Image Features from Scale-In variant Keypoints", *in International Journal of Computer Vision*, vol 60(2), pp. 91110, 2004.

[8] T.R. Tavares, G.G. Cabral, S.S Mattos, "Preprocessing unbalanced data using weighted support vector machines for prediction of heart disease in children", *The 2013 International Joint Conference on Neural Networks (IJCNN)*, Page(s): 1 - 8.

[9] I. Hamid,S. Mubarak, "Recognizing Complex Events Using Large Margin Joint Low-Level Event Model," *Lecture Notes in Computer Science Volume* 7575, 2012, pp 430-444.

[10] P. Gaspar, J. Carbonell, J.L. Oliveira, "On the parameter optimization of Support Vector Machines for binary classification", *Journal of Integrative Bioinformatics*, 9(3):201, 2012.

[11] O. Chapell, V. Sindhwani, S. S. Keerthi, "Optimization Techniques for Semi-Supervised Support Vector Machines", *Journal of Machine Learning Research*, pp203-233, 2008.

[12] Stephen Wright - Optimization Algorithms in Support Vector Machines, Computational Learning Workshop, Chicago, June 2009.

[13] C.A. Mitrea,I. Mironica, B. Ionescu, R. Dogaru, - Video Surveillance Classificationbased Multiple Instance Object Retrieval: Evaluation and Dataset. Int. Conf. on Intelligent Computer Communication and Processing, ISBN 978-1-4799-6568-7, pp. 171-179.

[14] R. Dogaru, A.T Murgan, S. Ortmann, M. Glesner, "A modified RBF neural network for efficient current-mode VLSI implementation," *Fifth International Conference on Microelectronics for Neural Networks*, pp. 265 - 270, 1996.

[15] R. Dograu, I. Dogaru, "An efficient finite precision RBF-M neural network architecture using support vectors," *Symposium on Neural Network Applications in Electrical Engineering (NEUREL)*, pp. 127 - 130, 2010.

[16] R. Genov and G. Cauwenberghs, "Kerneltron: support vector machine in silicon," *IEEE Trans. on Neural Networks*, vol. 14, no.5, pp. 1426-1434, 2003.

[17] T. Ojala, M. Pietikinen, D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," *in IAPR*, vol. 1, pp. 582 - 585, 1994.

[18] O. Ludwig, D. Delgado, V. Goncalves, U. Nunes, "Trainable Classifier Fusion Schemes: An Application To Pedestrian Detection," *IEEE Int. Conf. On Intelligent Transportation Systems*, vol. 1, pp. 432-437, 2009.

[19] J. Van De Weijer, C. Schmid, J. Verbeek, "Learning color names from real-world images", *in Computer Vision and Pattern Recognition*, 2007.

[20] M. Stricker, M. Orengo, "Similarity of color images", *in SPIE Conference on Storage and Retrieval for Image and Video Databases*, 1995.

[21] R. Dogaru, "A hardware oriented classifier with simple constructive training based on support vectors", *in Proceedings of CSCS-16, the 16th Int'l Conference on Control Systems and Computer Science*, Vol.1, pp. 415-418,2007.