A Fisher Kernel Approach for Multiple Instance Based Object Retrieval in Video Surveillance

Ionuț MIRONICĂ[1], Cătălin Alexandru MITREA[1], Bogdan IONESCU[1,2] , Patrick LAMBERT[2]
[1]*LAPI, University Politehnica of Bucharest, 061071 Romania*
[2]*LISTIC, University of Savoie Mont Blanc, Annecy, France*
imironica@imag.pub.ro

***Abstract*— This paper presents an automated surveillance system that exploits the Fisher Kernel representation in the context of multiple-instance object retrieval task. The proposed algorithm has the main purpose of tracking a list of persons in several video sources, using only few training examples. In the first step, the Fisher Kernel representation describes a set of features as the derivative with respect to the log-likelihood of the generative probability distribution that models the feature distribution. Then, we learn the generative probability distribution over all features extracted from a reduced set of relevant frames. The proposed approach shows significant improvements and we demonstrate that Fisher kernels are well suited for this task. We demonstrate the generality of our approach in terms of features by conducting an extensive evaluation with a broad range of keypoints features. Also, we evaluate our method on two standard video surveillance datasets attaining superior results comparing to state-of-the-art object recognition algorithms.**

***Index Terms*— automated video surveillance, Fisher Kernel representation, multiple-instance object retrieval.**

I. INTRODUCTION

During the last years, mainly because of the recent turbulent world events, the automated video surveillance techniques became an important research field. Fast developments in digital camera and video processing technology facilitated the availability of intelligent video surveillance systems basically in any public places. However, they provide only the infrastructure to capture, store and distribute video documents, while leaving the task of event detection mainly to human operators. Manually analyzing footage is a highly labor-intensive and time consuming task.

Today, a fully automated indexed video surveillance system is not commercially available. In the last years, most of the existing research progress was made for behavior, motion detection and human tracking methods. However, the main limitation of automated video surveillance remains in the searching capabilities. Once one has identified a possible target event, the system is not able to provide tracking capabilities of the entities causing that event during previous or future recordings, e.g., finding the other crimes where the burglar was involved in. Currently, this is actually done manually, by human operators. Considering the fact that a typical video surveillance system, in its simplest form (using only one video source), involves the recording of countless hours of footage, manually searching within records is hugely time consuming and at the same time inefficient and often unreliable. In practice, video surveillance systems feature tens of video sources, making the problem even more challenging. An automated surveillance system should help the operator in detecting certain persons, and make it possible to discover unlawful activities more quickly (either in real-time or by searching in existing video footage). The goal is to eventually have a system that can quickly and accurately monitor large and very complex areas for human behaviors, and when needed to report observed activities to an operator, or even deploy assistance if required.

The objective of this work is to discuss a solution for a system capable of providing content-based search capabilities within multiple-source video surveillance footage.

The proposed system is based on Fisher Kernel (FK) and Support Vector Machines (SVMs) and is capable of automatically identifying the occurrences of a certain person of interest during the video footage. After a human operator selects the person to search for from one of the frames, our system does the retrieval in two steps. Firstly, we extract the contour of persons by using a motion detection approach. For each contour that contains a specific person we compute a set of keypoints. Then, we train a Gaussian Mixture Model (GMM) with these features, and determine the Fisher Kernel representation with respect to this new GMM. Finally, a SVM is trained using the initial human feedback, yielding a specialized classifier in the new feature space.

This paper extends our previous work in [1] by introducing a new Fisher Kernel representation framework for video surveillance, including evaluation on new datasets and considering more feature extraction schemes. In [1] we propose a new relevance feedback algorithm based on Fisher Kernel representation in the context of multimodal video classification (using the visual, audio, motion and textual information). The algorithm is developed specifically for capturing in particular video temporal variation for video scenes/sequences classification. In contrast, the novel contributions of this work can be synthesized with the following:

- We propose a novel, frame-based, method for automated content-based retrieval of regions of interest in video surveillance that exploits a combination of Fisher Kernels and SVMs;
- We demonstrate the generality of our approach by evaluating it on a broad range of keypoint descriptors. We achieve better performance than other state-of-the-art approaches whereas evaluation is carried out on two

Part of this work was supported under InnoRESEARCH POSDRU/ 159/1.5/S/132395, ExcelDOC POS-DRU/159/1.5/S/132397 (2014-2015) and SCOUTER PN-IIIN- DPST-2012-1-0034 (2013-2015).

standard datasets [2, 3]. This makes the results both relevant and reproducible.

The reminder of the paper is organized as follows. Section II discusses several relevant video surveillance approaches and situates our work accordingly. The proposed system is presented in Section III. Section IV reports the experimental results. Finally, Section V provides a brief summary and concludes the paper.

II. Previous Work

Traditional passive video surveillance has two main drawbacks: (1) finding available human resources to observe the output is expensive and; (2) manual systems are ineffective when the number of cameras exceeds the ability of human operators to keep track of the evolving scene.

Currently, video surveillance systems are mostly passive. They require a human operator to monitor the video feeds on a screen, and to alert security crews when their assistance is required in case of emergency. In order to remove these drawbacks, over the recent years there have been extensive research activities in proposing new ideas, solutions and systems for robust automated surveillance systems. A large number of methods are reported in recent surveys [4]. In general, all existing approaches rely on efficient content description of the video information as an intermediate step, namely: color and texture [5], shape [6], audio [7] and feature points [8].

For instance, Landabaso et al. [9] introduced a robust multimodal tracking and classification system, that takes into account multiple characteristic features (e.g., velocity, shape, colour) of a 2D object appearance simultaneously in accordance with their respective variances. The system also further incorporates a classification module to classify each persistently tracked object, based on the analysis of local repetitive motion changes within the blob representation over a period of time. Ikizler-Cinbis and Sclaroff [13] extracted multiple features on the human, objects and scene, and employed a multiple-instance learning framework for human action recognition. Yang and Ramanan [14] proposed a method for articulated human detection and human pose estimation in videos based on a new representation of deformable part models. They detect small bounding boxes with a multi-scale Histograms of Oriented Gradients (HoG) descriptor, instead of complete body limbs, making their work more efficient because it prevents the problem of double counting. The body part detector combined with the Histograms of Optical Flow (HoF) features obtained good results on daily living activities [15]. However, this framework is adapted to a specific task and requires the use of motion compensation for foreground estimation and the detection and tracking of the human in the scene, generating a high computational cost. The accuracy of the algorithm is highly correlated with the performance of the human detector.

More recently, most of the contribution has been made to find automatic ways of describing video contents with parameters having enough representative power for the retrieval task. The approaches focused on the understanding of video contents using the visual and spatio-temporal information [10]. For instance, Muller-Schneiders et al. [11] proposed a real-time video surveillance system which was specifically designed for a low volume of false positives because surveillance guards might get deviated by too many alarms caused by, e.g., rain, trees, varying illumination conditions or small camera motion. This system uses the temporal information of the video, e.g., Cuboid detector, Hessian 3D detector or SURF 3D [12]. In spite of their good performance, feature descriptors are limited by their computational complexity (e.g., processing a large-scale video database may take days or weeks) that makes them unsuitable for real-time scenarios.

In this respect, current research addresses the development of low complexity algorithms to combine global with local strategies. One alternative is to use the Fisher Kernel representation. The Fisher Kernel theory was introduced by Jaakkola et al. [16] to combine generative and discriminative methods. Specifically, a collection of features is represented by its gradient with respect to a generative distribution. The resulting vector is then used in discriminative classifiers. Fisher Kernels were introduced in computer vision by Perronnin et al. [17], which applied the FK framework to represent collections of local visual features such as SIFT [8] using Gaussian Mixture Models as generative distribution. FKs found their application in other fields as well as, starting from web genre classification, event classification [9] to topic-based text segmentation [20] and web audio genre classification [19]. Aran and Akarun [18] introduced a multi-class classification strategy for a sign language data set. More recently, the Fisher representation was used by Myers et al. [21] for detection of user-defined events. They propose a set of multimodal features (i.e., audio, motion, visual) together with a set of late fusion techniques.

In this paper we adapt this particular class of methods for the design of an automated surveillance system. We introduce a new approach designed specifically for classification that uses a combination of Fisher Kernel representations and Support Vector Machines (SVM) classifiers. The FK representation has been successfully applied to many fields, but to the best of our knowledge, the FK have never been used in automated video surveillance. The FK representation is particularly suited for this scenario because it highlights the frames that contain occurrences of certain objects of interests. Experimental validation on two standard datasets proved the superiority of this approach compared to other state-of-the-art methods from the literature.

III. The proposed system

The proposed system works as in the following. The operator selects from few frames a region of interest of the object / human that needs to be searched in the database. Then, the system uses these frames to create a model for the searched object. This step defines the query to the system. Based on the user's interrogation, on the next layer, the system automatically searches in the entire database all the instances of the object / human to be found. The architecture of the proposed system is presented in Figure 1, and it consists of four different layers, namely:

(1) Firstly, the cameras collect the video information, which is transmitted to the motion detection layer. This module targets the extraction of moving objects, such as

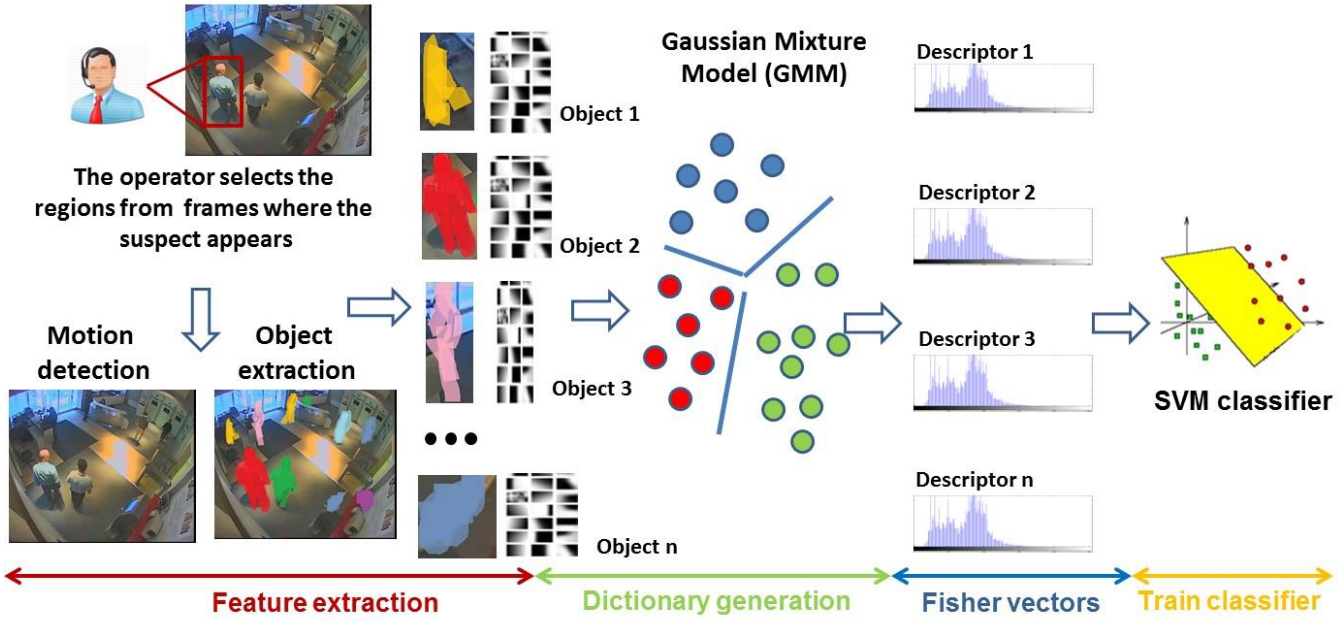

Figure 1. The proposed automated surveillance system: (1) motion detection and feature extraction, (2) dictionary representation, (3) Fisher Kernel representation, (4) SVM classification.

persons or cars. Motion analysis is very important because it optimizes the next stages performance by selecting relevant information, removing the irrelevant frame sections and so reducing the computational load. For each extracted object a set of keypoints are extracted. Feature extraction component addresses the creation of visual patterns for each segmented moving object in the video. We extracted the keypoints by using several state-of-the-art algorithms. These approaches were chosen due to their robustness, compact representation and significance for human perception;

(2) Having these keypoints, we learn a generative Gaussian Mixture Model [24] from the extracted keypoints;

(3) Then, we represent all the objects using a Fisher Kernel representation with respect to this GMM.

(4) The final step is represented by the discriminative training step, thus, we train a SVM classifier on the Fisher Kernel vectors. We apply this SVM and we obtain a final ranking.

A. *Motion detection*

Motion detection algorithms represent the first component of our system. These algorithms have as main purpose to obtain motion information, which further is required for objects' extraction. A widely-used technique for moving object segmentation is the background subtraction, which compares color or intensity of pixels in adjacent video frames. Significant differences are attributed to object motion. For this paper, we used the method presented in [25], where the authors propose the use of a Gaussian probabilistic density function (pdf) on the most recent *n* frames. Each pixel is characterized by mean μ_t and variance σ_t^2, and it is classified as object if the following condition is accomplished:

$$\frac{|(I_t - \mu_t)|}{\sigma_t} > th \qquad (3)$$

where I_t is the intensity of the current pixel, and the *th* represent a threshold (a common setting is to have *th* = *2.5*). We choose this method because it obtained good results in automated surveillance tasks [1] and proved robust to different types of noise and illumination changes.

B. *Feature extraction*

We extract a set of keypoints for each moving object. In order to describe the visual content, we compute the following features:

Scale-Invariant Feature Transform (SIFT) [8] represents a standard for the local image description. The computation of the SIFT descriptor consists of several steps. First, a set of orientation histograms is created on 4 x 4 pixel neighborhoods with 8 bins each. These histograms are computed from magnitude and orientation values of samples in a 16 x 16 region around the keypoint such that each histogram contains samples from a 4 x 4 sub-region of the original neighborhood region. The magnitudes are further weighted by a Gaussian function with standard deviation σ equal to one half the width of the descriptor window. Finally, the descriptor becomes a vector of all the values of these histograms.

Speeded Up Robust Features (SURF) [22] represents another robust local feature representation. SURF uses the sum of the Haar wavelet responses around the point of interest, which can be calculated very fast with the aid of the integral image.

Pyramid Histogram Of visual Words (PHOW) [23] are a variant of dense SIFT descriptors, extracted at multiple scales (e.g., 5, 7, 10, 12 pixels). It uses a color space version, named PHOW-color that extracts descriptors on the three HSV image channels.

In order to compute these features we used the VLFeat library [27], maintaining the default settings as provided in [26].

C. Fisher Kernel proposed approach

The main idea behind Fisher Kernel (FK) representation is to describe a signal as the gradient of the probability density function that is a learned generative model of that signal.

Intuitively, such representation measures how to modify the parameters of the probability density function in order to best fit the signal, similar to the measurements in a gradient descent algorithm for fitting a generative model [16]. The gradient vector is, by definition, the concatenation of the partial derivatives with respect to the model parameters. Let μ_i and σ_i be the mean and the standard deviation of *i*'s Gaussian centroid, $\gamma(i)$ be the soft assignment of descriptor μ_i to Gaussian *i* (with *t = 1, ..., T*), and let *D* denote the dimensionality of the descriptors x_t. $G^x_{\mu,\sigma,i}$ is the D-dimensional gradient with respect to the mean μ_i and standard deviation σ_i of Gaussian *i*. Mathematical derivation leads to [17]:

$$G^x_{\mu,i} = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^{T} \gamma(i) \frac{x_t - \mu_i}{\sigma_i} \quad (1)$$

$$G^x_{\sigma,i} = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^{T} \gamma(i) \left[\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right] \quad (2)$$

where the division between vectors is a term-by-term operation. Using this representation, the final gradient vector G_x, i.e., our new descriptor, is the concatenation of the $G^x_{\mu,i}$ and $G^x_{\sigma,i}$ vectors, for *i = 1, ..., T*. This leads to a $2 \cdot T \cdot D$, where *D* represents the size of keypoint features.

Initially, the user selects a few frames (approximately 2-3 seconds of video footage - *n* frames), where the sought person appears. Then, we add to these frames a set of unrelated *k* frames which will be used as negative examples. Using all these frames together, i.e., *n + k*, we train a Gaussian Mixture Model on the keypoints features. The GMM contains several parameters which impact the performance of the algorithm: the number of clusters *c*, the size of keypoints features and the normalization techniques. First of all, to make the Fisher Kernel computationally feasible, we apply PCA on the original keypoints vectors of the frames. After having obtained the mixture model, we convert the original features of the frames into the Fisher Kernel representation using Equations 1 and 2. The final step is represented by the use of normalization, applied on the final Fisher Kernels. We applied the L_2 and power normalization to the final vector.

D. Classification

The training step is represented by a two-class Support Vector Machine (SVM) classifier. The classic binary SVM training algorithm builds a linear margin that maximizes the distance between two classes. SVMs can efficiently perform a non-linear classification by using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The SVM approach is remarkably tolerant on the relative sizes of the number of training examples of the two classes. In our algorithm, the SVM model is trained on the n + k frames, according to the user's feedback. After a training step, all the documents are ranked according to the SVMs confidence level. At the end, a final ranking is obtained, by ordering the classifier's output confidence levels [28].

The SVM algorithm usually depends on several parameters. One of them, denoted *C*, controls the tradeoff between margin maximization and error minimization. Also, additional parameters may appear for non-linear mapping into feature space, namely the kernel parameters. In most of the experiments these parameters are globally tuned for the dataset [29]. However, a better strategy is to approximate the optimal value of these parameters at query level. In line with this, we divide the feedback samples in two parts: one for training, and one for the evaluation of the SVM parameters performance. We change the values of these parameters until the optimal parameters are obtained. This approach is not computational expensive mainly because the training and evaluation steps are done on a reduced set of results. We use two types of SVM kernels: a fast linear kernel and the RBF nonlinear kernel. While linear SVMs are very fast in both training and testing, SVMs with an RBF kernel are more accurate in many classification tasks.

IV. Experimental results

A. Datasets

The validation of the proposed approach was carried out on two standard video datasets, namely: Scouter [2] and PEVID-HD [3] (see Figure 2).

Scouter: represents an indexed video collection that contains several complex automated surveillance scenarios. It is composed by videos documents, acquired with several video surveillance cameras installed in the convention hall of UTI Grup company. The dataset consists of 30 video documents (3 different days x 10 cameras). The videos are recorded at 6 to 10 fps, with a resolution of 704 x 675 pixels. In total, the collection contains (3 days) x (10 cameras) x (average 120 seconds clip) x (10 frames per second) = approximately 36,000 annotated frames;

PEVID-HD: consists of 21 video clips (16 seconds each, full HD 1,920 x 1,080 pixels, 25 fps). Video clips show people performing various actions in indoor and outdoor environments during day and night times. The people shown in the videos are of different gender and ethnicity.

These datasets are in particular challenging due to the diversity of video footage, and specifically the variability of videos within the same categories. Also, the video footage contains variable lighting conditions as well as different levels of difficulty and includes several challenges such as noise, low quality image or blurring, increasing the difficulty of its analysis. Figure 2 illustrates some image examples in this respect.

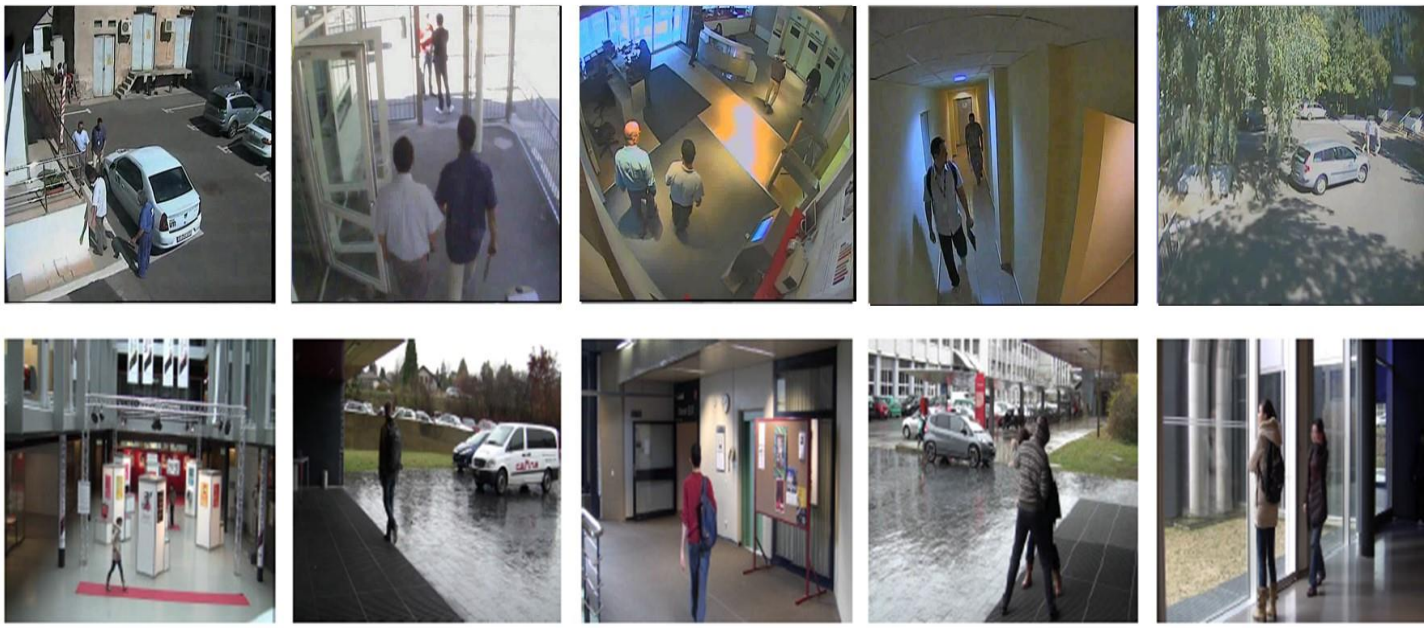

Figure 2. Sample frames from the Scouter [2] (first line) and PEVID-HD (second line) [3] datasets.

B. Evaluation

To assess retrieval performance, we use a global measure of performance, the Mean Average Precision (MAP), which is computed as the mean of the average precision scores for each query:

$$MAP = \sum_{q=1}^{Q} \frac{AP(q)}{Q} \tag{4}$$

where *Q* represents the number of queries, and *AP()* is:

$$AP = \frac{1}{m} \sum_{k=1}^{n} \frac{f_c(v_k)}{k} \tag{5}$$

where *n* is the number of frames, *m* is the number of frames of category *c*, and v_k is the *k-th* frame in the ranked list $\{v_1,...,v_n\}$. Finally, $f_c()$ is a function which returns the number of frames of category *c* in the first *k* frames if v_k is of category *c* and *0* otherwise (we used the trec_eval scoring tool available at http://trec.nist.gov/trec_eval/).

Also, we compute the classical precision and recall. Precision represents the proportion of the true positives against all the positive results (measure of false positives) and recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database (measure of false negatives). We also compute the F_β -Score [42] that combines the precision and the recall:

$$F_\beta = (1+\beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

where *b* represent the parameter that allows us to weigh recall more than precision or vice versa. This is an important property of the F_β Score, and is the primary reason why this measure was chosen. Recall from an automated surveillance system should produce no false-negatives and a minimal number of false-positives. For this reason, recall is weighted twice as much as precision by setting β = 2 when calculating F2-Score.

To validate our approach we conducted several experiments which are presented in the following. The first experiment (Section C) provide an experiment that studies the influence of motion detector on the algorithm's performance. The second experiment (Section D) motivates the choice of the best feature keypoints for the retrieval and we study the influence of Fisher Kernel parameters on systems accuracy. The third experiment (Section E) deals with comparing our method with other relevant algorithms from the literature. Finally, we provide a computational efficiency discussion on the proposed framework (Section F).

C. The evaluation of motion detectors

In this experiment we study the influence of motion detection algorithms on the system's performance. We tested three types of motion detectors: a background subtraction motion detector [25], an accumulative optical flow approach [40] and the Kalman filter motion detector in [41] (see Section III-B). In this experiment we tested only the performance of the motion detection with the objective of successfully retrieving the moving persons. Evaluation is performed by comparing the results to the actual ground truth. The performance of each motion detector is presented in Table I. Good accuracy is obtained with the Kalman filter motion detector, namely 81%. On the other hand, background subtraction motion detector obtains better performance, accuracy is equal to 87%. The lowest performance is obtained with the accumulative optical flow, which has been shown to be very sensitive to the parameter tuning.

TABLE I. COMPARISON OF MOTION DETECTORS ALGORITHMS.

Motion detection algorithm	**Accuracy**
Background subtraction motion	87%
Kalman filter motion detector	81%
Accumulative optical flow method	72%

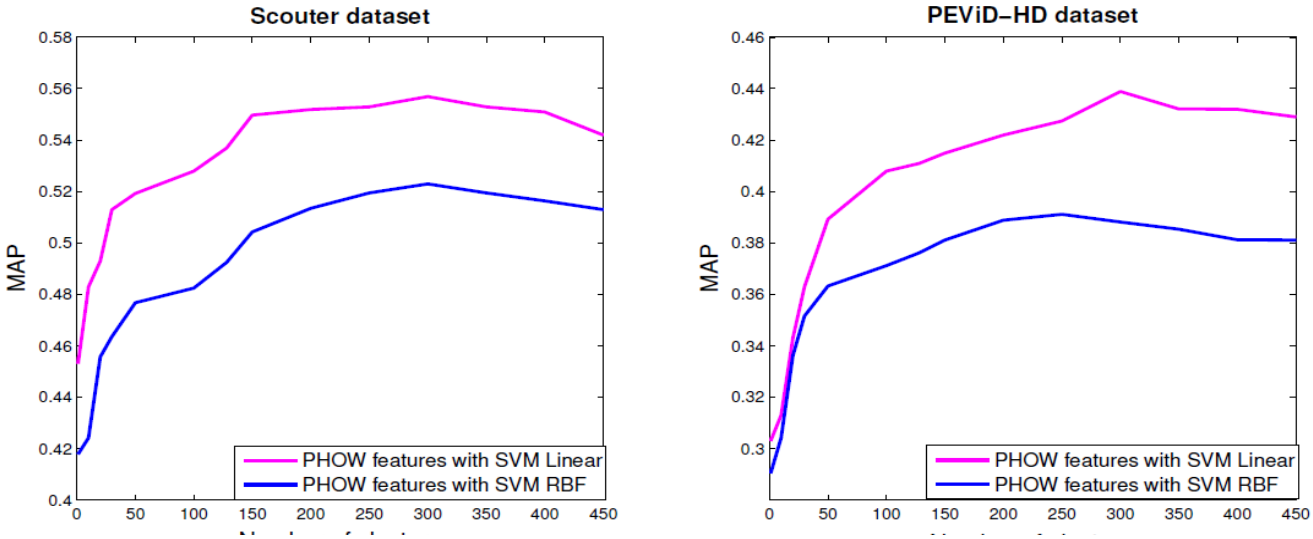

Figure 3. Mean Average Precision (MAP) while varying the number of cluster centers.

D. Parameter tuning

In this experiment we study the influence of Fisher Kernel parameters on the system's performance. First of all, to make the Fisher Kernel computationally feasible, we apply Principal Component Analysis (PCA) on these features and reduce the dimensions by around 40%. This step represents a common practice for many object recognition algorithms [17, 26, 30, 31]. We also applied the L_2 and power normalization, that was demonstrated to improve the performance of Fisher kernel [17].

The first parameter used in our approach is represented by the local descriptor algorithms, used for the description of the keypoints. We tested three different algorithms: SIFT, SURF and PHOW (which represent common features for image retrieval tasks and gives good results on Pascal VOC datasets [31]). The results are presented in Table II. For both datasets, the PHOW local descriptors have the highest stability and robustness: 55.69% MAP for the Scouter dataset and 43.21% MAP for the PEVID-HD dataset. SIFT also proves high stability in many situations, but it provide high sensitivity at illumination changes. Overall, the performance of SIFT is with 2-4% lower than the PHOW features. SURF provides faster speed but also it has many drawbacks, e.g., it is not stable to rotation and illumination changes. Therefore, it provides the lowest percentage of MAP values: 52.12% for the Scouter dataset and 38.22% for the PEVID-HD dataset.

TABLE II. COMPARISON BETWEEN SYSTEM ACCURACY (MAP, PRECISION, RECALL, F2SCORE VALUES) USING DIFFERENT KEYPOINTS ALGORITHMS.

Dataset	Evaluation parameter	**SURF**	**SIFT**	**PHOW**
Scouter Dataset	MAP	52.12%	53.67%	**55.69%**
	Precision	47.49%	48.56%	**50.21%**
	Recall	71.18%	72.21%	**74.11%**
	F2Score	64.72%	65.80%	**67.66%**
PEVID-HD	MAP	38.22%	39.57%	**43.21%**
	Precision	33.73%	34.21%	**37.47%**
	Recall	61.74%	62.19%	**67.83%**
	F2Score	52.94%	53.44%	**58.37%**

In the second experiment we analyze the influence of the number of centroids. The results are presented in Figure 3. One can observe that the performance increases with increasing the number of centroids. The best performance is obtained with 250 centroids. After this value, the performance decreases with 1 percent. A big improvement can be noticed compared to version with only one centroid: for the Scouter dataset the MAP parameter goes from 42% to 51% and from 45% to 55.69% for the SVM with RBF and Linear kernels. Also, for the PEVID-HD dataset the increase of number of centroids significantly improves the results: from 29% to 38% and from 31% to 43.21% for the SVM with RBF and Linear kernels.

The last parameter that has to be taken into consideration is the SVM kernel. The second experiment shows that we obtain better results with linear kernel.

E. Comparison with state-of-the-art

In order to compare our algorithm with other approaches, we have selected the settings that provide the greatest improvement in performance: 250 GMM centroids, PHOW features and SVM classifier with a linear kernel. The final experiment consists of comparing our approach with several state-of-the-art descriptors and classifiers pairs. Given the specificity of the task, i.e., automated video surveillance, we tested several visual descriptors which are known to perform well on image retrieval tasks, namely: Histograms of Oriented Gradients (HoG) features [15], Color Naming histograms (CN) [16], color moments (CM) [17], Local Binary Pattern (LBP) [18] and Bag of Words (BoW) [19] (with SIFT and PHOW features). Also, we train these features using a broad category of classifiers: nearest neighbor (KNN) [20], Random Forests (RF) [20], linear SVM and SVM with RBF and Chi-Square kernel classifier [14]. Figure 4 presents the values of MAP scores for other state-of-the-art algorithms. On the Scouter dataset, the best results from state-of-the-art is obtained by BoW (with PHOW features) using SVM with Chi-Square kernel, namely 49.26%. Similar performances are performed with HoG features with KNN classifier and Color Naming histograms with SVM Chi-Square (46.03% and 45.08%). On the other hand, the color moment features obtain lower MAP rates with 9 to 10 percents. Similar results are obtained for the PEVID-HD dataset: the BoW (on PHOW features) with SVM classifier obtain 32.41%, while color

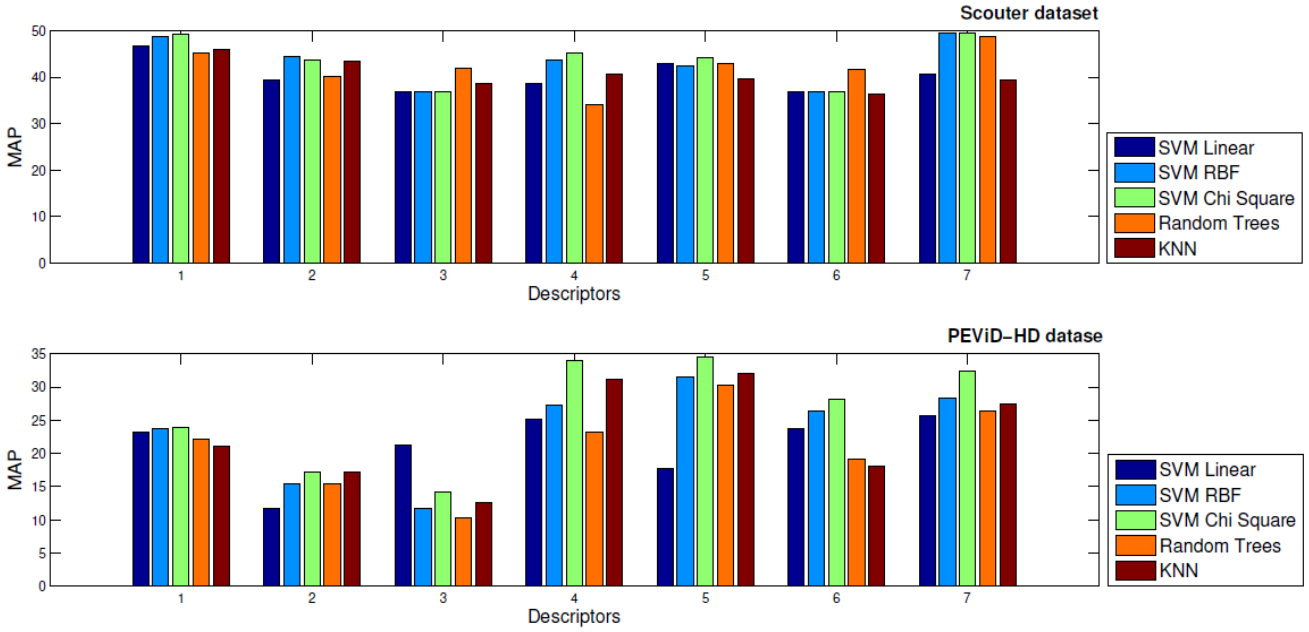

Figure 4. Mean average precision values using various descriptor - classifier combinations (1 - HoG, 2 - LBP, 3 - CM, 4 - CN3x3, 5 - CSD, 6 - BoW-SIFT, 7 - BoW-PHOW).

naming feature has a MAP value of 33.93%.

We present our results compared with the best state-of-the-art results in Table III. The proposed approach has the highest values for both datasets. We obtain 55.69% for the Scouter dataset and 43.21% for the PEVID-HD dataset. This represents an improvement of more than 6 percents for the Scouter dataset and more than 11 percents for the PEVID-HD dataset.

TABLE III. COMPARISON OF SYSTEM PERFORMANCE USING THE PROPOSED APPROACH AND THE BEST STATE-OF-THE-ART APPROACHES (MAP VALUES).

Descriptor	**Classifier**	**MAP**
Scouter dataset		
Fisher kernel with PHOW features	Linear SVM	**55.69%**
BOW with PHOW features	SVM with Chi-Square kernel	49.26%
HOG	KNN	46.03%
Color naming	SVM with Chi-Square kernel	45.08%
Baseline	Random decision	10.18%
PeVID-HD dataset		
Fisher kernel with PHOW features	Linear SVM	**43.21%**
Color naming	SVM with Chi-Square kernel	33.93%
BOW with PHOW features	SVM with Chi-Square kernel	32.41%
Baseline	Random decision	8.37%

We conclude that the proposed approach improves the retrieval performance, outperforming some other existing approaches, e.g., BoW, HoG, color naming, etc.

Figure 6 presents several system responses, when we use the best system configuration (Fisher Kernel with PHOW features and Linear SVM). The first query provide five examples of true positives (TP) examples in which the object found by the system are correctly identified according to the ground truth (note the scenario difficulty, different fields of view, object dimension, different object color, illumination, camera noise and other objects around the object of interest). Anyway there are also two false negatives situations (NT) in which the system is unable to classify correctly (according to ground truth) the object detected due to the signal noise, illumination conditions (insufficient, over exposed), partial object view (out of frame, junction with another object) or dimension too low. A similar example is also provided for the PEViD-HD dataset. Five of them represents frames that contains objects correctly identified according the ground truth. On the other side, we also provide several examples where the system is not able to provide correct response.

F. Computational complexity

In this section we discuss the computational complexity of the proposed description framework. We analyze the time for computing each processing step, from feature extraction to video classification. We perform this experiment on the Scouter dataset which contains more than 36,000 of video frames. The run-time is evaluated on a regular PC machine using a 2.9 GHz Intel Xeon CPU and 24GB of RAM. We do not use parallelization. Experiments were run with SIFT features and Linear SVM classifier. The computational cost per frame is presented in Figure 5. Descriptor extraction takes 150 milliseconds (ms) per image. The input/output operation lasts 30 ms per frame. The Fisher computation is very fast, namely 32 ms per frame. Finally, classification takes 8 ms for all classes.

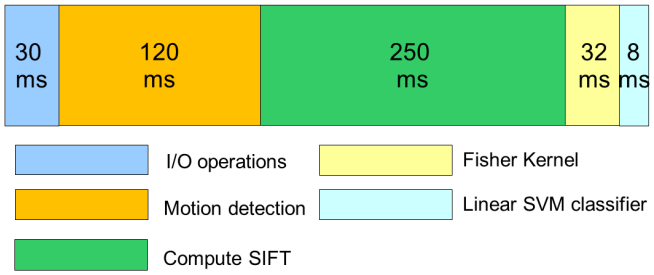

Figure 5. Total computational time (ms) per frame for the proposed video surveillance framework (Scouter [2] dataset).

A processing chain would take 450 ms per frame (12 seconds for 1 second of video, i.e., 25 frames). However, the most time-consuming components (i.e., motion detection and SIFT computation) can be computed only once, when the video footage is recorded. Therefore, we can take into consideration only the last two components which would take 40 ms per frame.

We conclude that this represents a reasonable, near real-time, cost considering the achieved performance. This is achieved without any algorithm optimization nor adequate hardware acceleration or parallel implementation. Using parallel processing will allow to easily achieve even better real-time performance.

V. CONCLUSIONS

In this paper we addressed the problem of content-based search for video surveillance. We formulated and analyzed a new approach that uses the Fisher Kernels theory. Our method consists of two steps: (1) altering the feature space by training a Gaussian Mixture Model on the reduced number of relevant frames and re-representing those features using Fisher Kernels; (2) a classification layer that uses a Support Vector Machine algorithm. We have tested several normalization techniques, keypoints features and discuss the influence of parameters on system's performance. Our experiments showed that our method always performs equally or better than other methods: Compared to the next best method, Bag of Words, we get an improvement on Scouter 6%, while for PEVID-HD we also get a higher improvement of 11% MAP. Also, we showed that we do not need large number of frames to train the FK framework, we achieve the best performance with only few examples. This makes the proposed approach implementable for a real time automated surveillance system.

Regarding the further continuation paths, future work will mainly consist in improving the computational speed of the proposed method. Also, we will adapt the method to address a higher diversity of video categories (use of the Internet) and we want to extend the Fisher kernel to other modalities, namely to use elaborated spatio-temporal features.

REFERENCES

[1] Ionuț Mironică, Bogdan Ionescu, Jasper Uijlings, Nicu Sebe, "Fisher Kernel based Relevance Feedback for Multimodal Video Retrieval", ACM International Conference on Multimedia Retrieval - ICMR 2013, Dallas, Texas, USA, April 16 - 19, 2013. [Online] http://dx.doi.org/10.1145/2461466.2461478

[2] C. Mitrea, I. Mironică, B. Ionescu, R. Dogaru, "Video Surveillance Classification-based Multiple Instance Object Retrieval: Evaluation and Dataset," International Conference on Intelligent Computer Communication and Processing (ICCP), ISBN 978-1-4799-6568-7, pp. 171-179, Cluj, Romania, 4-6, September, 2014. [Online]. Available: http://dx.doi.org/10.1109/ICCP.2014.6936970

[3] P. Korshunov, T. Ebrahimi, "PEViD: Privacy Evaluation Video Dataset Applications of Digital Image Processing," Proceedings of SPIE International Society for Optics and Photonics, vol. 8856, pp. 512-522, 2013. [Online]. Available: http://dx.doi.org/10.1117/12.2030974

[4] J. Aggarwal, J. Ryoo, "Human activity analysis: A review," In ACM Computing Surveys (CSUR), vol. 43(3), pp. 162-205, 2011. [Online]. Available: http://dx.doi.org/10.1145/1922649.1922653

[5] D. Duque, H. Santos, P. Cortez, "The OBSERVER: An Intelligent and Automated Video Surveillance System," In Proceedings of the International Conference on Image Analysis and Recognition (ICIAR), ISBN. 978-3-540-44893-8, pp. 989-909, 2006. [Online]. Available: http://dx.doi.org/10.1007/11867586_81

[6] Y. Mingqiang, K. Kidiyo, R. Joseph, "A Survey of Shape Feature Extraction Techniques," In International Conference of Pattern Recognition (ICPR), ISBN 978-953-7619-24-4, pp. 43-90, Tampa, Florida, USA, 8-11 December, 2008. [Online]. Available: http://dx.doi.org/10.5772/6237

[7] W. Choi, J. Rho, D. Han, H. Ko, "Selective background adaptation based abnormal acoustic event recognition for audio surveillance," In International IEEE Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 118-123, Beijing, China, 18-21 Sept. 2012. [Online]. Available: http://dx.doi.org/10.1109/AVSS.2012.65

[8] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," In International Journal of Computer Vision (IJCV), ISSN 0920-5691, vol. 60(2), pp. 91-110, 2004. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[9] J.L. Landabaso, X. L-Q. Xu, M. Pardas "Robust tracking and object classification towards automated video surveillance", In International Conference of Image Analysis and Recognition (IAR), ISSN 0302-9743, vol. 32(12), pp. 463-470, Porto, 29-30 September, 2004. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-30126-4_57

[10] B. Benfold, I. Reid, "Stable multi-target tracking in real-time surveillance video," In Computer Vision and Pattern Recognition (CVPR), pp. 3457-3464, Colorado Springs, USA, 21-23 June, 2011. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2011.5995667

[11] S. Muller-Schneiders, T. Jager, H. Loos, W. Niem, "Performance evaluation of a real time video surveillance system", In IEEE Intenational Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance(VS-PETS), ISBN 0-7803-9424-0, pp. 137-144, 2005. [Online]. Available: http://dx.doi.org/10.1109/VSPETS.2005.1570908

[12] J. Stottinger, B. T. Goras, N. Sebe, A. Hanbury, "Behavior and properties of spatio-temporal local features under visual transformations," In Proceedings of the International ACM Conference on Multimedia (ACM MM), ISBN: 978-1-60558-933-6, pp. 1155-1158, Florence, Italy, 25-29 October 2010. [Online]. Available: http://dx.doi.org/10.1145/1873951.1874174

[13] N. Ikizler-Cinbis, S. Sclaroff, "Object, scene and actions: combining multiple features for human action recognition," In Proceedings of the European Conference on Computer Vision (ECCV), vol. 6311, pp. 494-507, Heraklion, Crete, Greece, 5-10 September, 2011. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-15549-9_36.

[14] Y. Yang, D. Ramanan, "Articulated human detection with flexible mixtures of parts," In IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 35(12):2878-2890, 2013. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2012.261

[15] N. Rostamzadeh, G. Zen, I. Mironică, J.R.R. Uijlings, N. Sebe, "Daily Living Activities Recognition via Efficient High and Low Level Cues Combination and Fisher Kernel Representation," In IEEE International Conference on Image Analysis and Processing (ICIAP), ISSN 0302-9743, pp. 431-441, 2013. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-41181-6_44

[16] T. Jaakkola, D. Haussler, "Exploiting Generative Models in Discriminative Classifiers," In International Conference on Advances in Neural Information Processing Systems II, ISBN:0-262-11245-0, pp. 487-493, 1998.

[17] F. Perronnin, J. Sanchez, T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification," In European Conference on Computer Vision (ECCV), LNCS 6314, pp. 143-156, 5-11 September, Heraklion, Crete, Greece, 2010. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-15561-1_11

[18] O. Aran, L. Akarun, "A Multi-Class Classification Strategy for Fisher Scores: Application to Signer Independent Sign Language Recognition," In Pattern Recognition, 43(5):1776-1788, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2009.12.002

[19] P.J. Moreno, R. Rifkin, "Using the Fisher Kernel Method for Web Audio Classification," In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ISSN 1520-6149, vol. 6, pp. 2417-2420, 5-9 June, 2000, Istanbul, Turkey. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2000.859329

[20] Q. Sun, R. Li, D. Luo, W. Xihong, "Text Segmentation with LDA-based Fisher Kernel," In Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, 2008. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2000.859329

[21] G. K. Myers, C. G. Snoek, R. Nallapati, J. van Hout, S. Pancoast, R. Nevatia, C. Sun, "Evaluating Multimedia Features and Fusion for Example-based Event Detection, " In International Journal of Machine Vision and Applications (MVAP), 25(1):17-32, 2014. [Online]. Available: http://dx.doi.org/10.1007/s00138-013-0527-8

[22] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, "Speeded-up robust features (SURF)," In Computer Vision and Image Understanding

(CVIU), vol. 110(3), pp. 346-359, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2007.09.014

[23] A. Bosch, A. Zisserman, X. Munoz, "Image classifcation using random forests and ferns," In IEEE International Conference on Computer Vision (ICCV), pp. 1-8, Rio de Janeiro, Brasil, 14-21 Oct. 2007. . [Online]. Available: http://dx.doi.org/10.1109/ICCV.2007.4409066

[24] C. M. Bishop, "Pattern recognition and machine learning," In New York: Springer, ISBN 978-0-387-31073-2, vol. 4, nr. 4, 2006.

[25] C. R. Wren, A. Azarbayejani, T. Darrell, A. P. Pentland, "Pfinder: real-time tracking of the human body," In IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), ISSN 0162-8828, vol. 19(7), pp.780-785, 1997. [Online]. Available: http://dx.doi.org/10.1109/AFGR.1996.557243

[26] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," In International Proceedings of British Machine Vision Conference (BMVC), pp. 1-12, Dundee, 29 August - 2 September 2011. [Online]. Available: http://dx.doi.org/10.5244/C.25.76

[27] A. Vedaldi, B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," In Proceedings of the International Conference on Multimedia (ACM MM), ISBN: 978-1-60558-933-6, pp. 1469-1472, 2008, http://www.vlfeat.org/.

[28] V.N. Vapnik, "Statistical Learning Theory," New York: John Wiley & Sons, ISBN: 978-0-471-03003-4, 1998.

[29] O. Chapelle, "Training a Support Vector Machine in the Primal," In Neural Computation, MIT Press, vol. 19(5), pp. 1155-1178, 2007. [Online]. Available: http://dx.doi.org/10.1162/neco.2007.19.5.1155

[30] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, J. C. van Gemert, J. R. R. Uijlings, J. He, X. Li, I. Everts, V. Nedovic, M. van Liempt, R. van Balen, F. Yan, M. A. Tahir, K. Mikolajczyk, J. Kittler, M. de Rijke, J.-M. Geusebroek, T. Gevers, M. Worring, A. W. M. Smeulders, D. C. Koelma, "The MediaMill TRECVID 2008 semantic video search engine," in Proceedings of the 6th TRECVID Workshop, Gaithersburg, USA, November 2008.

[31] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results", http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[32] O. Ludwig, D. Delgado, V. Goncalves, U. Nunes, "Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection," In IEEE International Conference On Intelligent Transportation Systems, vol. 1, pp. 432-437, St. Louis, USA, 4-7 October, 2009. [Online]. Available: http://dx.doi.org/10.1109/ITSC.2009.5309700

[33] J. van DeWeijer, C. Schmid, J. Verbeek, D. Larlus, "Learning color names for real-world applications," in IEEE Transactions on Image

[34] Processing, ISSN 1057-7149, vol. 18(7), pp. 1512-1523, 2009. [Online]. Available: http://dx.doi.org/10.1109/TIP.2009.2019809

[35] M. A. Stricker, M. Orengo,"Similarity of color images," In Symposium on Electronic Imaging: Science and Technology, vol. 2420, pp. 381-392, 1995. [Online]. Available: http://dx.doi.org/10.1117/12.205308

[36] T. Ojala, M. Pietikinen, D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," In International Conference on Pattern Recognition (IAPR), vol. 1, pp. 582 - 585, Jerusalem, Israel, 09-13 Oct 1994.

[37] J. R. R. Uijlings, A. W. M. Smeulders, R. J. H. Scha, "Real-Time Visual Concept Classification," In IEEE Transactions on Multimedia, ISSN: 1520-9210, vol. 12(17), pp. 665-681, 2010. [Online]. Available: http://dx.doi.org/10.1109/ICPR.1994.576366.

[38] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, "When Is Nearest Neighbor Meaningful?", Database Theory ICDT Lecture Notes in Computer Science, ISSN 0302-9743, vol. 1540, pp. 217-235, Jerusalem, Israel, 10–12 January, 1999. [Online]. Available: http://dx.doi.org/10.1007/3-540-49257-7_15.

[39] L. Breiman, "Random forests," In Journal of Machine Learning, 45(1),2009. [Online]. http://dx.doi.org /10.1023/A:1010933404324

[40] N. Lu, J. Wang, L. Yang, Q. H. Wu, Motion Detection Based On Accumulative Optical Flow and Double Background Filtering, in World Congress on Engineering, pp. 602-607, 2007. [Online]. Available: http://dx.doi.org/10.1007/11596981_43.

[41] A. Bovik, The essential guide to video processing, Elsevier Inc, 2009, ISBN: 0123744563.

[42] G. Hripcsak, A. Rothschild, Agreement, the f-measure, and reliability in information retrieval, in Journal of the American Medical Informatics Association, vol. 12(3), pp. 296-298, 2005. [Online]. Available: http://dx.doi.org/10.1145/2009916.2009971.

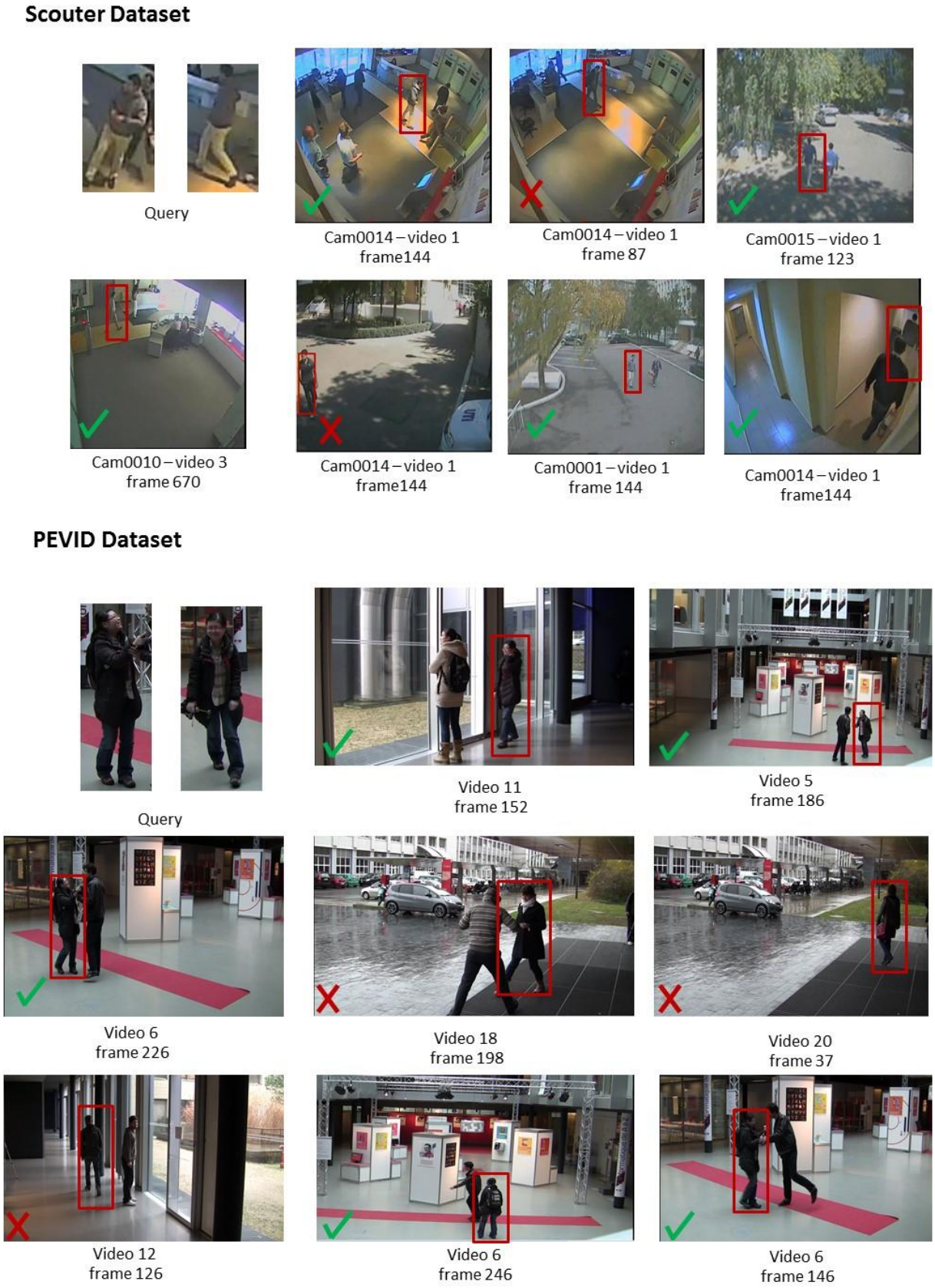

Figure 6: First two images represents the query. The retrieved results are marked with the red rectangles - ranking order from left (highest) to right. Correct detections are denoted by green (ok) whereas false detections are depicted with a red x.