Noname manuscript No.
(will be inserted by the editor)

Active Query Process for Digital Video Surveillance Forensic Applications

Dalia Coppi · Simone Calderara · Rita Cucchiara

Received: date / Accepted: date

Abstract Multimedia forensics is a new emerging discipline regarding the analysis and exploitation of digital data as support for investigation to extract probative elements. Among them, visual data about people and people activities, extracted from videos in an efficient way, are becoming day by day more appealing for forensics, due to the availability of large video-surveillance footage. Thus many research studies and prototypes investigate the analysis of soft biometrics data, such as people appearance and people trajectories. In this work we propose new solutions for querying and retrieving visual data in an interactive and active fashion for soft biometrics in forensics. The innovative proposal joins the capability of transductive learning for semi-supervised search by similarity, and a typical multimedia methodology based on user-guided relevance feedback to allow an active interaction with the visual data of people, appearance and trajectory in large surveillance areas. Approaches proposed are very general and can be exploited independently by the surveillance setting and the type of video analytic tools.

Keywords Multimedia forensics · Transductive learning · Relevance feedback

1 Introduction

Rainer et al. [2] in a recent paper discuss the difference between multimedia forensics and computer forensics, pointing out that the latter is a broad branch concerning the general involvement of computer in crime activities, while the former discipline is more specific and regards the extraction, from different sensors (cameras, microphones, etc.), of digital data that can be probative elements in many investigations. As an extension, we call video surveillance multimedia forensic the subset of activities

D.I.E.F. - University of Modena and Reggio Emilia
Via Vignolese 905/b, 41125 Modena, Italy
Tel.: +39 059 2056270
Fax: +39 059 2056129
E-mail: name.surname@unimore.it

carried out to extract probative elements from the massive quantity of surveillance videos now available to investigators.
Video surveillance data are often unreliable due to many factors, such as poor color resolutions, low frame-rate, occluded viewpoints, bad luminance. Working with these uncertain data, the experience of investigators becomes essential; the continuous knowledge transfer between users and machines and the role of human operator in evaluating query results are central and the user's deduction and feedback are invaluable elements that concur to gain a continuous improvement and refinement of automatic results. Most of the applications devoted to digital surveillance forensics allow to perform queries on some specific people related data, but so far the user involvement in the search process has been limited limited to executing subsequent queries over the obtained results. A typical system off line classifies and clusters similar data, and user iteratively work on pre-elaborated data. This approach can be referred to as **iterative passive querying** and its framework is shown in Fig. 1. The user, according to the capability of a GUI, can iterate the search and explore the results in a predefined result-space where only a limited number of results is returned. Any Google-like interface is an example belonging to this category of retrieval systems.
Conversely, in new-generation multimedia applications, user feedbacks are collected and adopted to improve the single query step, [11]. Precisely, the user has the ability of re-rank displayed results suggesting new positive and negative elements then used in the following retrieval step. After this manual refinement, the system, must be able to improve the query results to better encounter the desires of the users. In contrast to the previous setting we refer to this approach as **iterative active querying** process, Fig. 1.
In the context of video surveillance forensics, where a considerable amount of data is available at time of investigation, the experience of the human operator is an invaluable element to improve the retrieval of probative elements. We propose to adopt the approach of relevance feedback, as defined for CBIR systems, also in the field of multimedia forensics. Given the high amount of biometric data and the vast number of similarity measures developed by the research community, we focus on people appearance and trajectories, which can be extracted in a sufficient reliable way through video surveillance and video analytic systems.
We aim to present a link between semi supervised classification and user relevance feedback to provide a general and mathematically well-founded methodology to include user feedbacks in the forensics query process in order to improve the search results exploiting the user's notion of correctness on the automatically retrieved results. To this purpose we propose to exploit transductive learning, a well-known approach in machine learning community that recently gained attention even in multimedia community [16], with the final purpose of incorporating such an inference element in the forensic query process.
The advantages of our proposal can be summarised in few key points:

– *flexibility*: the proposed method can be adopted on any available feature similarity measure;

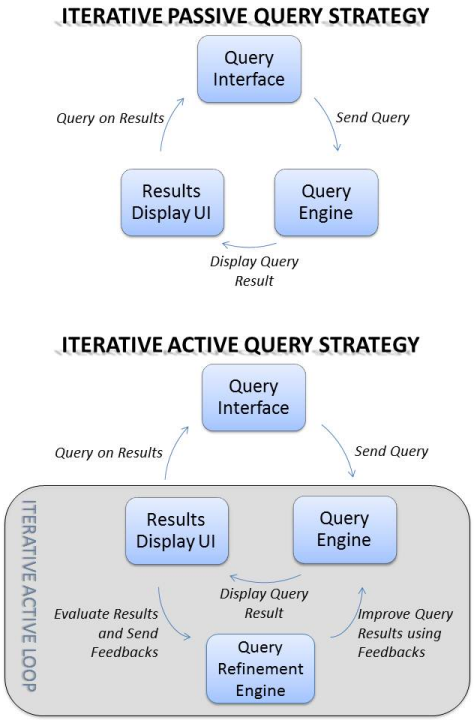

Fig. 1 Iterative Passive Query Scheme where no user feedback is provided in contrast to the Iterative Active Query Scheme where user feedbacks directly affect the query results.

- *interactivity*: user can operate directly on classification results in order to iteratively improve the precision of the system;
- *mathematical consistence*: the method is strongly related to other semi supervised learning method and graph search algorithms. There are some strong connections also with Markovian Networks and Bayesian Random Walks;
- *performance*: with some optimization strategies we can apply this solution to thousands of images in order to improve conventional nearest neighbour queries.

In this paper we initially discuss the use of transductive learning as a powerful tool for providing relevance feedback in search; then we propose our system which provides interactive active querying on forensics data and in particular on people appearance and trajectories.
The main novelty of our proposal consists in the use of a well known technique of semi-supervised learning together with the user relevance feedback, as intended for CBIR, in an iterative loop. We demonstrate how these two techniques can jointly improve the recall with respect to other methods.

2 Overview of semi supervised learning

What we aim to propose is a solution for seamlessly including user selection in query process in order to retrieve or discard elements following the user suggestions. The problem can thus be formulated as finding a technique for classifying the dataset according to the similarity with a given query where constraints on the results are imposed by the investigator using a visual interface. These settings can be reinterpreted as the question of learning a dichotomic classifier able to separate elements in query and non-query results using user imposed labels on the dataset and consequently classifying unlabelled elements.
Semi-supervised learning (SSL) refers to a subset of learning techniques that exploits both labelled and unlabelled data for training a classifier. SSL methods are placed halfway between supervised learning, where only labelled samples are used, and unsupervised learning, where samples are all unlabelled.
In the recent years SSL has been used in many areas including computer vision where several applications i.e. image segmentation [8] or object recognition and tracking [21] are re-interpreted as a semi-supervised classification problem. The reason for this choice can be explained observing that in many real world application is relatively easy to acquire a large amount of unlabelled (unclassified) data while it is quite difficult and expensive to have labelled, or classified, data. For example, in digital forensic tasks, to support investigations, images and videos can be obtained from surveillance cameras but their corresponding classification require a slow human annotation. Therefore, being able to utilize the unlabelled data is desirable to avoid the bottleneck constituted by the initial labelling of training data. Moreover the advantage of a semi-supervised approach is not only based on the fact of being computationally cheaper but also on its practical value in learning faster, better and its capacity to solve any problem otherwise solvable with supervised learning.

2.1 Semi-supervised learning methods

The general configuration of a SSL algorithm provides a data set $X = (x_i)_{i\in\{1,\dots,n\}}$ that can be divided in two parts: the elements $X_l = \{x_1, \dots, x_l\{$ with associated labels $Y_i = \{y_1, \dots, y_l\}$, and the elements $X_u = \{x_{l+1}, \dots, x_{l+u}\}$ whose labels are learned by the classifier.
Beyond this common basic configuration the state-of-the-art is composed by several different semi-supervised learning methods. Following [17] they can be classified in Generative Models, SSL Support Vector Machines (SVM), Co-Training and Graph-Based Models.

Generative models are probably the oldest semi-supervised learning method. They assume the form of a joint probability $p(x, y|\theta) = p(y|\theta)p(x|y, \theta)$, where the class prior distribution $p(y|\theta)$ can be a multinomial over Y and the class conditional distribution $p(x|y, \theta)$ can be an identifiable mixture distribution on X, for example a multivariate Gaussian. A baseline way to solve the problem consists of maximizing

the joint probability, that can be done by direct gradient descent or more conveniently by applying the Expectation Maximization (EM) algorithm. **Support Vector Machines** build a structure on the complete sample set, including both the labelled and unlabelled data, and assume that the decision boundary $f(x) = 0$ is situated in a low-density region (in terms of unlabelled data) between the two classes $y \in \{-1, +1\}$, aiming to find the hyperplane that separates the training data with the largest margin. In **Co-Training** the idea is to make use of different views on the objects to be classified and multiple, different learners are trained on the same labelled data, each classifier is specialized to a particular view and, since the true label is missing, all the learners must agree on the unlabelled data. Recently the most active area of research in semi-supervised learning is represented by **Graph-Based** methods. In this class of methods data are denoted by the nodes of a graph $G = (V, E)$, where the nodes V are the labelled and unlabelled instances and the undirected edges E connect instances i, j with a weight proportional to the similarity among them. The final aim of graph based learning methods is to predict labels for the unobserved nodes.
A significant distinction in SSL is the separation in **inductive** and **transductive** methods.
In the former category the learner uses both labelled training data $\{(x_i, y_i)\}_{i=1}^{l}$ and unlabelled data $\{x_i\}_{i=l+1}^{l+u}$ to synthesize a prediction function $f : X \mapsto Y, f \in F$ where F is the hypothesis space. The goal is to find a predictor able to classify future test data better than a predictor learned from only labelled data. Conversely **transductive learning** is solely interested in the predictions on the unlabelled training data $\{x_i\}_{i=l+1}^{l+u}$ without any intention to derive a generalized function for future test data. If label predictions are only required for a given test set, transduction can be argued to be more direct than induction: while an inductive method infers a function $f : X \mapsto Y$ on the entire space X, and afterwards returns the evaluations $f(x_i)$ at the test points, transduction consists in directly estimating the finite set of test labels [20].
This is in accordance with the transductive inference setting introduced by Vapnik [19] whose theory is based on the principle: *when trying to solve some problem, one should not solve a more difficult problem as an intermediate step*.

Semi Supervised Learning in video surveillance

In surveillance forensics settings semi-supervised methods can help the investigations in the task of processing the large amount of available digital videos. Usually only a small fraction of the information encoded in these videos is relevant to the investigators. Digital video surveillance can help to speed up the process of evaluating the recorded data in terms of extracting potentially useful information, e.g. information regarding when a target person entered or left the scene, the track of his movements or the analysis of his trajectory. These automatically extracted data can be examined by investigators using content based retrieval systems and performing queries by similarity on specific elements, SSL can thus be exploited to account for users knowledge about the query considering user feedbacks as labelled elements and the dataset as unlabelled elements. At this aim graph based SSL are intrinsically transductive, i.e.

they return only the value of the decision function on the unlabelled points and not the decision function itself.

3 Transduction on graphs using spectral theory

In this section we recall the general configuration of a graph based transductive learner and explain how to utilize this Semi Supervised Learning technique to retrieve relevant elements from a dataset of surveillance data exploiting user's feedbacks as labelled elements. The Transductive Learning (TL) configuration we propose here uses both positive and negative labelled elements and was previously introduced by Joachims [12].
Suppose to have a set of labelled instances $X_l = \{(x_i, y_i)\}_{i=1}^{l}$ where x_i are the input elements described by their features, e.g. people patch or people trajectory, and $y_i = \pm 1$ the corresponding labels. The labels assume the value $y_i = +1$ ($y_i = -1$) according respectively to user's positive (negative) feedbacks on the results retrieved by the system at the previous iteration. Suppose also to have a set of unlabelled instances $X_u = \{x_i\}_{i=l+1}^{l+u}$ which are the other elements in the dataset. The complete dataset comprises both the model X_l and the candidates samples X_u:

$$D(X, Y) = \{X_l \cup X_u, Y : y_i = \pm 1 \text{ iff } x_i \in X_l\} \tag{1}$$

By setting the problem in this form we aim at propagating, at each iteration, the knowledge transferred by the user' feedbacks into the labelled models that can be equivalently interpreted as the problem of estimating the missing label values in order to classify the complete dataset on the basis of labelled instances and the similarity among the elements. With these hypothesis the TL algorithm is the straightforward way to solve the problem.
Given an undirected graph $G = (V, E)$ with V the set of nodes representing elements in X and E the edges representing the similarity among them the adjacency matrix of the graph is A, whose elements a_{ij} are obtained from nodes distances $\rho(x_i, x_j)$ between samples x_i and x_j by an inverse exponential smoothing function with fixed bandwidth σ^2:

$$a_{ij} = \exp\left(-\frac{\rho(x_i, x_j)}{\sigma^2}\right) \tag{2}$$

the objective of the TL algorithm is to find a cut of the graph that separates positive and negative elements X_+ and X_-, or equivalently to find labels for unlabeled elements. Considering D as the diagonal degree matrix $D_{ii} = \sum_j A_{ji}$, and according to [14], we can compute the Laplacian graph as $L = D - A$, or in its normalized version $L = D^{-1}(D - A)$, and the TL solves the problem by solving the following minimization problem:

$$\begin{aligned} &\min_{\overrightarrow{z}} \overrightarrow{z}^T L \overrightarrow{z} + c(\overrightarrow{z} - \gamma)^T I(\overrightarrow{z} - \overrightarrow{\gamma}) \text{s.t.} \\ &\text{s.t. } \overrightarrow{z}^T \overrightarrow{z} = n \text{ and } \overrightarrow{z}^T 1 = 0 \end{aligned} \tag{3}$$

Where $\overrightarrow{z}$ is the generalized partition vector with elements z_i, $\overrightarrow{\gamma}$ equals $\gamma_+ = \sqrt{\frac{|\{i : z_i < 0\}|}{|\{i : z_i > 0\}|}}$ if $i \in X_+$ and $\gamma_- = -\sqrt{\frac{|\{i : z_i > 0\}|}{|\{i : z_i < 0\}|}}$ if $i \in X_-$ and c is a parameter that trades off training errors versus cut value.

Stepping back, the minimization problem in Eq. 3 can be obtained starting from the assumptions that the corresponding inductive learner should have low leave-one-out error (a) and constraining the problem to have averages over examples with similar expected value in the training and in the test set (b). The assumption (a) can easily be solved minimizing the LOO error on classification using a trivial KNN. The LOO error of the classifier can be bounded by:

$$Err_{loo}^{knn}(X, Y) \leq \sum_{i=1}^{N}(1 - \delta_i) \tag{4}$$

where δ_i is the KNN margin $\delta_i = y_i \frac{\sum_{j \in KNN(x_i)} y_j a_{ij}}{\sum_{m \in KNN(x_i)} a_{im}}$ with a_{ij} the similarity between x_i and x_j. The minimization of Eq. 4 can be obtained by maximizing the margin δ_i and imposing constrained values on the model labels, margin maximization can be written in matrix form leading to the following constrained optimization problem:

$$\begin{aligned} &\max_y y^T A y \text{ s.t.} \\ &y_i = \pm 1 \text{ if } x_i \in X_l \\ &\forall y_{j \neq i} \in \{-1, 1\} \end{aligned} \tag{5}$$

Even if this problem can be efficiently solved using both the s-t Mincut algorithm [1] or transductive SVM it usually leads to unbalanced cuts. The assumption (b) is therefore necessary to avoid this issue and the cut size can be accounted using a ratio-cut algorithm [10]. The traditional ratio-cut is an unsupervised problem and find the optimal solution is known to be NP hard. The constraint on y makes the problem semi-supervised, however letting y to assume real values and exploiting spectral properties of the graph Laplacians yields to an efficient way to find a solution to the balanced ratio-cut problem in a semi-supervised way. The ratio-cut problem minimizes the average weight of the cut leading to balanced a cut of the graph.

$$\begin{aligned} &\max_y \frac{cut(G^+, G^-)}{\mid \{i : y_i = 1\} \mid\mid \{i : y_i = -1\} \mid} \text{ s.t.} \\ &y_i = 1 \text{ if } i \in Y_l \text{and positive} \\ &y_i = -1 \text{ if } i \in Y_l \text{and negative} \\ &\overrightarrow{y} \in \{+1, -1\}^n \end{aligned} \tag{6}$$

Given the Laplacian L of the graph and the partition vector $\overrightarrow{z}$ the ratio-cut becomes

$$\min_{\overrightarrow{z}} \frac{\overrightarrow{z}^T L \overrightarrow{z}}{\overrightarrow{z}^T \overrightarrow{z}} \text{ with } z_i \in \{\gamma_+, \gamma_-\} \tag{7}$$

where γ_+ and γ_- are defined as previously. Even this problem is NP complex its relaxed version is solved exploiting the *Courant-Fischer Minimax Principle* stating that the second eigenvalue of L is the non-degenerate solution and the corresponding eigenvector, i.e. the Fielder Vector, solves the *argmin* problem. The unconstrained ratio-cut problem is unsupervised, therefore in order to take into account labelled instances a quadratic penalty can be introduced to the objective function in Eq. 7 obtaining Eq. 3. The optimization problem in Eq. 3 can be recast as a *Quadratic Eigenvalue Problem*, (QEP) and solved analytically for positive semi definite matrices using eigendecomposition [18]. Specifically, given the eigendecomposition $L = U \Sigma U^T$ of the Laplacian, and introducing $\overrightarrow{w} = U^{-1} \overrightarrow{z}$, the constraint in Eq. 3 becomes equivalent to setting $w_1 = 0$ because the eigenvector of the smallest eigenvalue is always $\overrightarrow{1}$. Redefining V and Λ as the matrices containing, respectively, all eigenvectors U and eigenvalues Σ except the smallest one, Eq. 3 can be rewritten as

$$\begin{aligned} &\min_{\overrightarrow{w}} \overrightarrow{w}^T \Lambda \overrightarrow{w} + c(V\overrightarrow{w} - \gamma)^T I (V\overrightarrow{w} - \overrightarrow{\gamma}) \text{ s.t.} \\ &\overrightarrow{w}^T \overrightarrow{w} = n \end{aligned} \tag{8}$$

Finally introducing $G = (\Lambda + cV^T V)$ and $\overrightarrow{b} = cV^T C \overrightarrow{\gamma}$ the objective function can be one more time rewritten, disregarding continuos terms, as $\overrightarrow{w}^T G \overrightarrow{w} - 2\overrightarrow{b}^T \overrightarrow{w}$. Following again the *Courant-Fischer Minimax Principle* the minimization in Eq. 8 is then solved for $\overrightarrow{w}^* = (G - \lambda^* I)^{-1} \overrightarrow{b}$ where λ^* is the smallest eigenvalue of

$$\begin{bmatrix} G & -I \\ -\frac{1}{n} \overrightarrow{b} \overrightarrow{b}^T & G \end{bmatrix} \tag{9}$$

I is the identity matrix. The optimal value of Eq.3 is computed as

$$\overrightarrow{z}^* = Ev \overrightarrow{w}^* \tag{10}$$

producing a predicted value for each example in the test set. The hard class assignment can be easily obtained thresholding the prediction vector $\overrightarrow{z}^*$.

4 Anatomy of an iterative active query system for people biometries retrieval

As mentioned in Sec. 1, in digital video surveillance the importance of the user feedbacks on the output obtained from a query is relevant because of its capacity of improving the results and better encounter the investigators purposes.
We aim at combining the proposed transductive learning approach with relevance feedback, where the positive and negative feedbacks given by the user iteratively constitute the labelled models of the transductive algorithm.
Queries simply based on similarity among objects often show poor quality and do not offer an effective tool for improving the search and achieving useful information. The

basic idea is to take advantage of the user feedback on the results retrieved from the first query and iteratively, and incrementally, use the modified set as the input labelled models of the transducer, Fig. 2. At every iteration irrelevant samples are potentially moved away from the query center, while far away relevant samples, not considered at the beginning, are attracted toward the query center. This leads to more accurate and precise results allowing an effective support to investigations.

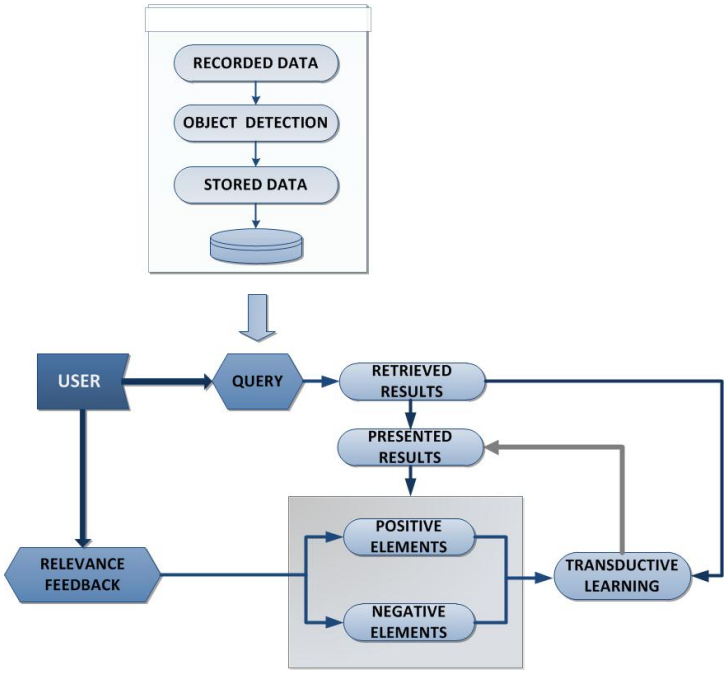

Fig. 2 Overview of the Iterative Active Query System where user can interact with presented results in order to feed the transductive learner with positive and negative examples.

The main steps of our system are the following:

1. Definition of a distance metric.
2. Choice of a query image q.
3. Query-by-example performed and a set of N results retrieved. The results are the first N Nearest Neighbour elements ranked with increasing distance from the query centre in the chosen feature space.
4. User gives feedbacks on the n presented results (with $n < N$). Elements are marked as relevant or not relevant i.e. positive or negative feedbacks
5. Affinity matrix (Eq. 2) computed over labelled and unlabelled elements in the set of N retrieved results.
6. Transductive learning performed and predicted value for each element in the test set computed with Eq. 10.

$\Longrightarrow$ Steps from 4. to 6. are repeated until, with subsequent refinement, the final result is reached, for a maximum of 5 iterations.

When processing video surveillance data several elements are interesting and can be automatically acquired by modern video surveillance systems. Among these people trajectories and people appearances constitute a proper choice that carry important information about people behaviour in the scene. Investigators may want to find people that follows specific paths and eventually searching for all the visual occurrences of a suspect in a video stream. From a technical point of view snapshots can be queried based on their visual similarity that embodies both colour and textural information, trajectories can be compared either on their position in the scene, **spatial analysis**, or their shape, **shape analysis**. We propose three different similarity measure for respectively people trajectories based either on their shape or points coordinates and for people snapshots. The measures can be used to perform queries by example and to build the affinity matrix of Eq. 2 depending on the queried feature.

4.1 Trajectories models for people paths analysis

The people trajectory projected on the ground plane is a very compact representation based on a sequence of 2D data coordinates $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, often associated with the motion status, e.g. the punctual velocity or acceleration. When data are acquired in a real system they should be properly modelled to account for tracking errors, noise in the support point extraction and inaccuracies; in particular, the modelling choice must take into account that when comparing two points belonging to different trajectories, small spatial shifts may occur and trajectories never exactly overlap point-to-point.

4.1.1 Trajectory spatial analysis

In our forensic application we adopted the spatial model proposed in [3], that combines a *statistical* representation of the data with a *point-to-point* approach to balance the computation cost and the accuracy. Briefly, given the k^{th} rectified trajectory projected on the ground plane $T_k = \{\mathbf{t_{1,k}} \dots \mathbf{t_{n_k,k}}\}$, where $\mathbf{t_{i,k}} = (x_{i,k}, y_{i,k})$ with n_k the number of points of trajectory T_k, a bivariate Gaussian $S_i^k = \mathcal{N}(x, y \mid \boldsymbol{\mu_{i,k}}, \boldsymbol{\Sigma})$ is centred on each data point $\mathbf{t_{i,k}}$ (i.e., having the mean equal to the point coordinates $\boldsymbol{\mu_{i,k}} = (x_{i,k}, y_{i,k})$) and with fixed covariance matrix $\boldsymbol{\Sigma}$. Using a sequence of Gaussians, one for every point, allows to build an envelope around the trajectory itself, obtaining a slight invariance against spatial shifts. After assigning a Gaussian to every trajectory point, the trajectory can be modelled as a sequence of symbols corresponding to Gaussian distributions $\overline{T}_j = \left\{S_{1,j}, S_{2,j}, ..., S_{n_j,j}\right\}$, where every symbol $S_{i,j}$ is associated to its bi-variate Gaussian.

4.1.2 Trajectory shape analysis

A completely different perspective of analysis consists in discarding the paths position in the scene focusing instead on their shape. This approach is oriented on discerning and synthesizing common and frequent motion patterns that are important indicators of people habits and interactions. Recently it has been proposed to perform the shape analysis in the directions domain, considering the trajectories as a sequence of angles and adopting a circular-defined statistic for modelling periodic angular data. Angular analysis seems a promising way to approach the shape comparison problem since sequences of angles are by definition location independent. We implemented the statistical model for shape analysis in [4], that exploits circular statistics to robustly model data points. In analogy with spatial model we aim at obtaining a sequence of symbols that statistically represent the sequence of angles that constitute the trajectory. Consequently, for handling angular data, circular distributions have been proposed in literature and among these the von Mises distribution demonstrates to be the most suitable since it is circularly defined and correctly capture the periodic nature of angular data, [4]. Von Mises distribution is thus an ideal pdf to describe a trajectory T_j by means of its angles. However, in the general case a trajectory is not composed only of a single main direction, thus it should be represented by a multi-modal pdf, and thus the model consists of a mixture of von Mises (MovM) distributions:

$$p\left(\theta\right) = \sum_{k=1}^{K} \pi_k \mathcal{V}\left(\theta|\theta_{0,k}, m_k\right) \tag{11}$$

where $\mathcal{V}\left(\theta|\theta_{0,k}, m_k\right) = \dfrac{1}{2\pi I_0(m)} e^{m\cos\left(\theta-\theta_0\right)}$ and with I_0 the modified Bessel function of order 0. When the mixture distributions components have to be learnt the EM algorithm is a natural solution to compute the parameters. In the case of MovM a full derivation of the EM has been proposed in [4]. Once the K components of the mixture are computed, every direction $\theta_{i,j}$ is encoded with a symbol $S_{i,j}$ with a MAP approach, where every symbol corresponds to the most probable MovM components.

4.1.3 Trajectory alignment based similarity measure

Despite the adopted model, since trajectories do not share always the same length, unless if re-sampled, we propose to adopt an alignment based similarity measure, properly designed either for spatial model or shape model. In the case of statistical models, it has been suggested in [3] that the well-known Needleman-Wunsh (NW) alignment algorithm is effective in comparing sequences of pdf while in [7] the efficiency of the Dynamic Time Warping(DTW) alignment algorithm in presence of very large datasets has been demonstrated. Independently from the chosen method, either NW or DTW, basically the alignment is performed by using dynamic programming with a computational complexity of $O(n*m)$, where m and n are sequences lengths, and exploiting specific recurrent relations, depending on the alignment algorithm, after the definition of a symbol-to-symbol distance measure.

In the case of symbol sequences that represent spatial-Gaussian probability distributions, a proper symbol-to-symbol similarity measure must be defined in order to perform the global alignment. Among the possible metrics to compare probability distributions we chose to employ the Bhattacharyya coefficient, to measure the distance between the two normal distributions $S_{a,k}$ and $S_{b,m}$ corresponding to a^{th} and b^{th} symbols of sequences $\overline{T}_k$ and $\overline{T}_m$, respectively:

$$\begin{aligned}\rho\left(\mathcal{N}_{a,k},\mathcal{N}_{b,m}\right) &= d_{BH}\left(\mathcal{N}\left(x,y|\boldsymbol{\mu}_{a,k},\boldsymbol{\Sigma}_a\right),\mathcal{N}\left(x,y|\boldsymbol{\mu}_{b,m},\boldsymbol{\Sigma}_b\right)\right)\\ &= \frac{1}{8}\left(\boldsymbol{\mu}_{a,k}-\boldsymbol{\mu}_{b,m}\right)^T\left(\overline{\boldsymbol{\Sigma}}\right)^{-1}\left(\boldsymbol{\mu}_a-\boldsymbol{\mu}_b\right)+\\ &\quad+\frac{1}{2}\ln\left(\frac{\det\overline{\boldsymbol{\Sigma}}}{\sqrt{\det\boldsymbol{\Sigma}_a\det\boldsymbol{\Sigma}_b}}\right)\end{aligned} \tag{12}$$

where $2\cdot\overline{\boldsymbol{\Sigma}}=\boldsymbol{\Sigma}_a+\boldsymbol{\Sigma}_b$. Since in our case $\boldsymbol{\Sigma}_a=\boldsymbol{\Sigma}_b=\boldsymbol{\Sigma}$, we can rewrite the distance as:

$$\rho\left(\mathcal{N}_a^k,\mathcal{N}_b^m\right)=\frac{1}{8}\left(\boldsymbol{\mu}_a-\boldsymbol{\mu}_b\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_a-\boldsymbol{\mu}_b\right) \tag{13}$$

When the data sequences is modeled, using the Mixture of Von Mises Model in Section 4.1.2, since the pdfs parameters space is not Euclidean, one possible symbol-to-symbol similarity measure between the univariate pdf associated to each symbol is the Bhattacharyya coefficient between Von Mises distribution,[13]. We can derive the score for the Mixture of Von Mises Model; specifically, we measured the distance between distributions p and q using the Bhattacharyya coefficient:

$$\rho\left(p,q\right)=\int_{-\infty}^{+\infty}\sqrt{p\left(\theta\right)q\left(\theta\right)}d\theta \tag{14}$$

following the derivation in [4] for two univariate Von Mises distribution the analytic form of the coefficient results:

$$\begin{aligned}&\rho\left(T_i,T_j\right)=\rho\left(\mathcal{V}\left(\theta|\theta_{0,i},m_i\right),\mathcal{V}\left(\theta|\theta_{0,j},m_j\right)\right)=\\ &\left(\sqrt{\frac{1}{I_0\left(m_a\right)I_0\left(m_b\right)}}I_0\left(\frac{\sqrt{m_i^2+m_j^2+2m_im_j\cos\left(\theta_{0,i}-\theta_{0,j}\right)}}{2}\right)\right)\end{aligned} \tag{15}$$

where m_k is the precision parameter of the distribution (inverse of variance), and, where it holds that $0\leq\rho\left(S_i,T_j\right)\leq 1$.

4.2 People Snapshots model

Given the similarity measure on people trajectories, people paths can be compared on the basis of common coordinates properties or shape. In conjunction with path information, video surveillance system can automatically extract an additional important cue about people identity, the visual appearance. Investigators, during the analysis of video data, may want to search for people visually similar to a selected reference

individual. For adding visual appearance capabilities to the proposed forensic application we implemented a snapshot similarity measure based on covariance matrix features descriptor, [15]. The covariance matrix is a square symmetric matrix $d \times d$, with d the number of selected features independently from the size of the image window, carrying the advantage of being a low dimensional data representation. Given the covariance matrix C its diagonal entries represent the variance of each feature and the non-diagonal entries represent the correlations; generally a single matrix extracted from a region is enough to match the region in different views and poses since the noise corrupting individual samples is largely filtered out with the average filter during covariance computation. Moreover covariance matrices can have scale and rotation invariance property and are independent to the mean changes such as identical shifting of color values. Let I be a three-dimensional color image and F be the $W \times H \times d$ dimensional feature image extracted from I,

$$F(x, y) = \Phi(I, x, y) \tag{16}$$

where the function Φ can be any mapping such as intensity, color, gradients, filter responses, etc. Let $\{z_i\}_{i=1...N}$ be the d-dimensional feature points inside F, with $N = W \times H$. The image I is represented with the $d \times d$ covariance matrix of the feature points:

$$C_R = \frac{1}{N-1} \sum_{i=1}^{n} (z_i - \mu)(z_i - \mu)^T \tag{17}$$

where μ is the vector of the means of the corresponding features for the points within the region R.
In our case z_i is the feature vector composed for each pixel by its spatial, color and edge information. We use x and y pixel location in the image grid, RGB color values and Gx and Gy first order derivatives of the intensities calculated through Sobel operator w.r.t. x and y. Therefore each pixel of the image is mapped to a seven-dimensional feature vector $z_i = [\, x \;\; y \;\; R \;\; G \;\; B \;\; Gx \;\; Gy \,]^T$. Based on this features vector the covariance of a region is a 7×7 matrix.

4.2.1 People snapshot similarity measure

To obtain the most similar region to the given object, we need to compute the distances between the covariance matrices corresponding to the target object and the candidate regions. However, the covariance matrices do not lie on the Euclidean space, therefore an arithmetic subtraction of two matrices would not measure the distance of the corresponding regions. The distance metric between the covariance matrices is proposed in [9] as the sum of the squared logarithms of the generalized eigenvalues.

$$\rho(C_i, C_j) = \sqrt{\sum_{i=1}^{d} ln^2 \lambda_k(C_i, C_j)} \tag{18}$$

where $\lambda_k(C_i, C_j)_{k=1...d}$ are the generalized eigenvalues of C_i and C_j computed as:

$$\lambda_k C_1 x_k - C_j x_k = 0 \;\; k = 0 \ldots d \tag{19}$$

where x_k are the generalized eigenvectors. The distance measure ρ satisfies the metric axioms, positivity, symmetry, triangle inequality, for positive definite symmetric matrices.

The three presented similarity measures can proficiently be injected in the query system to perform query by appearances, selecting and filtering people trajectories and finally constructing affinity matrices, Eq. 2, for applying transduction based on user feedbacks.

5 Experiments on people soft biometries retrieval

To evaluate the impact of relevance feedback on the querying system precision and recall we collected two different types of soft biometric data, namely people trajectories and people snapshots, from publicly available datasets. Trajectories have been acquired from Edinburgh Informatics Forum Pedestrian Database [1]. This dataset contains several days of people trajectories taken from a bird-eye view camera. Additionally, snapshots have been collected from the publicly available CAVIAR dataset [2] extracting them from videos using a conventional HOG based people detector, [6]. The dataset for quantitative accuracy evaluation consists of 4000 trajectories and 3000 people snapshots manually ground-truthed. In all the tests a query element was selected by the user and the first 30 results returned by the system. The baseline query method is the Nearest Neighbour classifier where the results are ranked according to the similarity w.r.t. the query element. We propose to use a naive KNN algorithm as baseline classification because of its simplicity of implementation: it does not require a training step, it can run using only one of few examples of the object of interest and the only input it needs is the similarity matrix computed among all elements. Besides these implementation reasons the other major advantage of the KNN is that, being based only on the similarity matrix, avoids issues that arises with other classifier (for example SVMs), using features that do not lie in the Euclidean space like covariance matrix, which require a Dissimilarity Space Embedding (DSE), or trajectories, that would require equivalently DSE or an alignent based kernel.
For every query a maximum of 5 iteration of the relevance feedback procedure of Sec. 4 are performed by the user that can select either positive feedbacks or both positive and negative examples and the improved ranking result is given by the label function of Eq. 10. The number of results retrieved by the NN classifier was chosen according to the average number of results presented in the first page of most of the search engines; while the maximum number of iteration was set empirically.
The tests have been performed evaluating the relevance feedback importance in three different types of query:

- Trajectory points query where we adopt a similarity measure for comparing people trajectories based on both their shape and coordinates in the image plane;

[1] http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING
[2] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1

– Trajectory shapes query where trajectories are compared using a measure acting directly on their shape by comparing the sequence of directions(angles) that compose the trajectories;
– Snapshot appearance similarity query where given the image of a subject similar images are then returned on the basis of both color and textural elements.

Fig. 3 underlines the average results on 100 queries where the user could freely select positive and negative examples. The boost on performances obtained through relevance feedback (bars portions over the dashed lines) is evident and demonstrates the capability of the system to obtain satisfying results even when simple and fast similarity measures are employed to compare the elements. It is remarkable to note that the final average precision and recall are closer, in most of the considered cases, to 90% even when exploiting a reduced number of iterations of the transductive classifier.

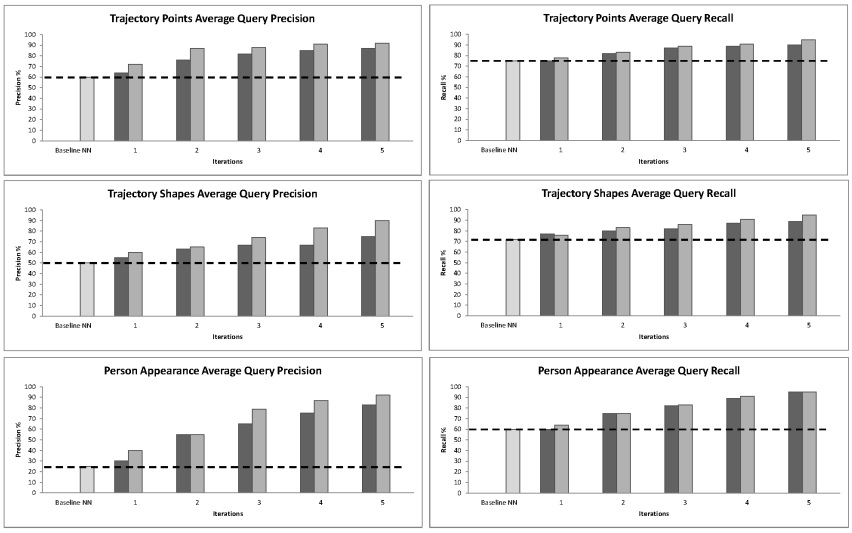

Fig. 3 Queries average precision and recall on three different kinds of features. Light grey bars refer to transduction with only positive feedbacks while dark grey bars refer to transduction with both positive and negative feedbacks. First bars and dashed lines represent the values of precision and recall of the trivial KNN baseline method.

We finally report a comparison between our methods and some other algorithms for relevance feedback originally introduced for CBIR. The algorithms are the following:

– Baseline KNN classifier: no multiple iterations of relevance feedback, the system proposes the firs k results according retrieved by the classifier;
– Naive relevance feedback (actually no relevance feedback at all): the system discards the current set of k results and proposes to the user the next k, following the original rank given by the visual similarity;

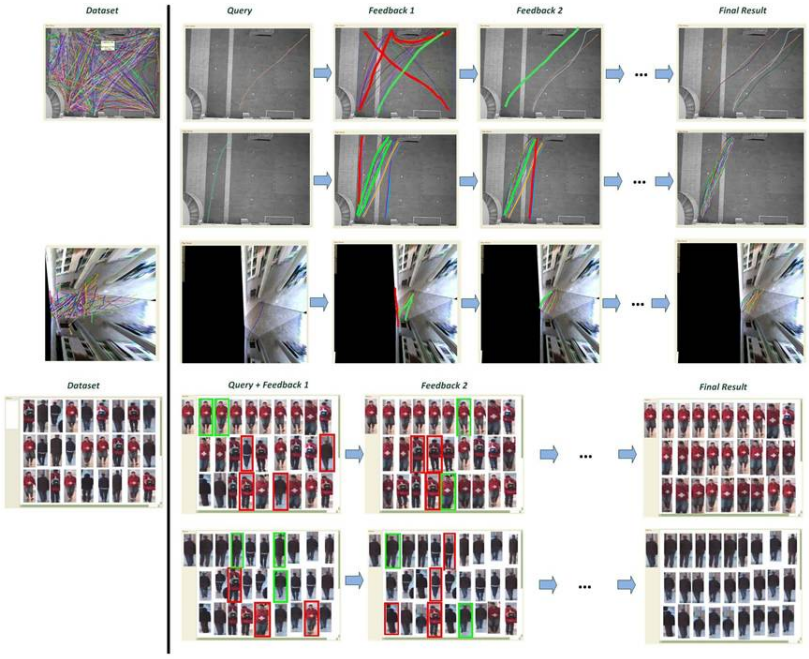

Fig. 4 Examples of query performed on our active query system for video surveillance forensic analysis. First two queries are performed on people trajectories and subsequent iteration, involving user feedbacks, are depicted. Green elements are positive selected feedbacks while red are the negative ones. Final results are shown on the rightmost sides of the rows. Last two queries are performed on people appearances similarity with the purpose of finding a specific subject.

Table 1 Recall values at different iteration steps.

method	1	2	3	4	5
Baseline	68	-	-	-	-
Naive	68,4	72,1	75,2	78,1	81,7
MSFSW	69,2	75,3	76,9	79,7	82,4
TLP	70,7	79,0	83,7	88,3	91,3
TLPN	72,7	80,2	86,2	91,1	95,0

– Original Mean Shift Feature Space Warping,MSFSW, proposed by [5], where the feature space or the metric space are manipulated, in order to shape it in the direction of the users' feedbacks;
– Transductive Learning with Positive feedbacks, TLP: our transductive learning approach which uses positive feedbacks as labelled samples;
– Transductive Learning with Positive and Negative feedbacks, TLPN: our transductive learning approach with both positive and negative feedbacks.

Performances have been evaluated in terms of incremental recall at each step, with a fixed number of 5 iterations, and results presented in Tab. 1 are a mean of the values obtained on people snapshots, trajectory points and shapes.

The comparison shows how the approach we proposed reaches an higher value of recall within the same number of iterations.
The developed active query engine has been integrated in our automatic video surveillance framework that is capable to automatically extract people trajectories and snapshots from static video surveillance cameras. Examples of queries results using subsequent active iteration by transduction, on both publicly available data and data from a real experimentation carried out from cameras installed in Modena, are depicted in Fig. 4. The figure shows on each row, different queries, on both people snapshots and trajectories, on the results retrieved in the first iteration the user can give his feedbacks marking in green or red respectively positive or negative results. The first query is performed using trajectories shape, while the second and the third one also exploits trajectories points. It is clear how bad results, are progressively moved away from the presented results and how the recall of the system iteratively increases.

6 Conclusions

In conclusion we presented a transductive learning setting to easily incorporate user feedbacks in the query process as an aid for forensic investigations. The proposed solution is general and applies to whichever distance measure defined for comparing elements by their similarity. Results are encouraging and demonstrate the effectiveness of both transduction and relevance feedback that can play a key role in forensic investigation process allowing an active approach that directly involves investigators knowledge and intuition in order to improve the results of automatic query systems. Currently Municipality of Modena (Italy) is testing the efficacy of the system in the investigative process, this experimentation aims at demonstrating the concrete advancement w.r.t. available commercial products that follow passive query scheme, while investigative process is by definition an active process that strongly relies on user interaction, for producing satisfying results for the stakeholders.

References

1. A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *In Intl. Conf. on Machine Learning (ICML)*, 2001.
2. R. Böhme, F. C. Freiling, T. Gloe, and M. Kirchner. Multimedia forensics is not computer forensics. In *Proceedings of the 3rd International Workshop on Computational Forensics*, IWCF '09, pages 90–103, 2009.
3. S. Calderara, A. Prati, and R. Cucchiara. Video surveillance and multimedia forensics: an application to trajectory analysis. In *Proc. of the ACM workshop on Multimedia in forensics, mifor*, pages 13–18, 2009.
4. S. Calderara, A. Prati, and R. Cucchiara. Mixtures of von mises distributions for people trajectory shape analysis. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(4):457 –471, 2011.
5. Y. Chang, K. Kamataki, and T. Chen. Mean shift feature space warping for relevance feedback. In *Proceedings of the 16th IEEE international conference on Image processing*, ICIP'09, pages 1829–1832, Piscataway, NJ, USA, 2009. IEEE Press.
6. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005.

7. H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. J. Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *In Proc. of VLDB*, 1:1542–1552, 2008.
8. O. Duchenne, J.Y. Audibert, R. Keriven, J. Ponce, and F. Segonne. Segmentation by transduction. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, pages 1–8, 2008.
9. W. Forstner, B. Boudewijn Moonen, and C.F. Gauss. A metric for covariance matrices, 1999.
10. L. Hagen and A.B. Kahng. New spectral methods for ratio cut partitioning and clustering. *Trans. Comp.-Aided Des. Integ. Cir. Sys.*, 11(9):1074–1085, November 2006.
11. F. Hopfgartner, D. Vallet, M. Halvey, and J. Jose. Search trails using user feedback to improve video search. In *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, pages 339–348, 2008.
12. T. Joachims. Transductive learning via spectral graph partitioning. In *In Intl. Conf. on Machine Learning (ICML)*, pages 290–297, 2003.
13. T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *Communication Technology, IEEE Transactions on*, 15(1):52 –60, february 1967.
14. U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.
15. F. Porikli, O. Tuzel, and P. Meer. Covariance tracking using model update based on lie algebra. In *Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 728–735, 2005.
16. H. Sahbi, J.-Y. Audibert, and R. Keriven. Graph-cut transducers for relevance feedback in content based image retrieval. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1 –8, 2007.
17. C. Sammut and G. I. Webb, editors. *Encyclopedia of Machine Learning*. Springer, 2010.
18. F. Tisseur and K. Meerbergen. The quadratic eigenvalue problem. *SIAM Review*, 43(2), 2001.
19. V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.
20. V. Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics) 2nd Edition*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
21. Y. Zha, Y. Yang, and D. Bi. Graph-based transductive learning for robust visual tracking. *Pattern Recognition*, 43(1):187–196, 2010.