Multi-Target Tracking by Online Learning of Non-linear Motion Patterns and Robust Appearance Models

Bo Yang and Ram Nevatia
Institute for Robotics and Intelligent Systems, University of Southern California
Los Angeles, CA 90089, USA
{yangbo|nevatia}@usc.edu

Abstract
We describe an online approach to learn non-linear motion patterns and robust appearance models for multi-target tracking in a tracklet association framework. Unlike most previous approaches that use linear motion methods only, we online build a non-linear motion map to better explain direction changes and produce more robust motion affinities between tracklets. Moreover, based on the incremental learned entry/exit map, a multiple instance learning method is devised to produce strong appearance models for tracking; positive sample pairs are collected from different tracklets so that training samples have high diversity. Finally, using online learned moving groups, a tracklet completion process is introduced to deal with tracklets not reaching entry/exit points. We evaluate our approach on three public data sets, and show significant improvements compared with state-of-art methods.

1. Introduction
Multi-target tracking in real scenes is important for many applications, such as surveillance, robotics, and human-computer interactions. This problem becomes difficult in crowded scenes with frequent occlusions and interactions among the targets.

In recent years, improvement in target detection has brought a trend of association based tracking approaches, which associate corresponding detection responses or tracklets (track fragments) into longer tracks [12, 16, 5]. Affinities between tracklets, i.e., the linking probabilities, are often evaluated as

\[ P_{\text{link}}(T_i, T_j) = A_m(T_i, T_j)A_a(T_i, T_j)A_t(T_i, T_j) \] (1)

where \( A_m(\cdot) \), \( A_a(\cdot) \), and \( A_t(\cdot) \) indicate the motion, appearance and temporal affinities between tracklets \( T_i \) and \( T_j \). A Hungarian algorithm is often used to find the global optimum [12, 6]. Though much progress has been made, motion affinity estimation and appearance modeling remain key issues that limit performance.

A linear motion model is commonly assumed for each target [18, 16, 3, 6]. However, as shown in Figure 1(a), there are often several non-linear motion patterns in a scene. Appearance models are often pre-defined [18, 16] or online learned from a few neighboring frames [6, 15]; tracklets with long gaps are difficult to be associated due to appearance changes.

Fortunately, there is often useful knowledge in the scene, such as motion patterns, entry/exit points, and moving groups, as shown in Figure 1, for solving the above problems. We describe an online learning method which automatically finds dynamic non-linear motion patterns and robust appearance models to improve tracking performance.

The framework of our approach is shown in Figure 2. Similar to [5], detection responses are first linked into tracklets by a low level association approach. Based on confident tracklets\(^1\), we online learn a non-linear motion map, which is a set of non-linear motion patterns, e.g., the orange tracklet in Figure 2, used for explaining non-linear gaps between other tracklets, e.g., the gap between two blue tracklets in Figure 2. For efficiency purpose, our tracking is done in sliding windows one by one; the non-linear motion map is updated by re-learning for each sliding window, as motion patterns may change with time.

Meanwhile, an online learning process is adopted to automatically find entry or exit points in a scene, as shown in green masks in Figure 2. We limit our approach to static cameras, where entry/exit points do not change in different sliding windows and can be learned incrementally. As shown in Figure 1(b), entry/exit points constrain the starting and ending of the trajectories; a tracklet ending before reaching exit points should be associated with other tracklets.

\(^1\)The definition is given in Section 3.1.
lets. We collect detection responses from different tracklets to form potential positive and negative bags, which are used in a multiple instance learning approach for building robust appearance models, so that tracklets tend to be associated until they reach entry/exit points, even under long gaps.

The Hungarian algorithm is utilized to find the global optimum; for the temporal affinity in Equ. 1, we use the same function as in [6]. For tracklets not reaching entry/exit points, we try to find whether they belong to any moving groups as shown in Figure 1(c). The trajectories of observed targets in the same groups, e.g., the green tracklet in Figure 2, are used to complete those of occluded unobserved ones, e.g., the pink tracklet in Figure 2, so that the latter can reach the entry/exit points.

The contributions of this paper are:
- An online learned non-linear motion map for explaining reasonable non-linear motions, as well as a new motion affinity estimation method.
- An incremental learning approach for an entry/exit map, and a multiple instance learning (MIL) algorithm for finding robust appearance models.
- An algorithm for automatically finding moving groups, which are further used for track completion.

The rest of this paper is organized as follows: related work is discussed Section 2; learning of the motion map for motion affinities is given in Section 3; Section 4 describes the entry/exit map estimation and the MIL algorithm for appearance models; learning moving groups for tracklets completion is presented in Section 5; experiments are shown in Section 6, followed by conclusion in Section 7.

2. Related work

Multi-target tracking has been a popular topic for several years. Visual tracking approaches often track each object separately [7, 21], but have difficulties to deal with large number of objects in crowded scenes. Association based approaches often optimize multiple targets globally and simultaneously [12, 16, 3, 20], and therefore are more robust to long time occlusions and crowded scenes.

Motion patterns have attracted researchers’ attention in recent years, and are often learned as priors for estimation of each target’s trajectory [4, 17, 13, 22]. It is often assumed that targets follow a similar motion pattern; trajectories not following the pattern are penalized. This assumption works well for extremely crowded scenes, where it is difficult for a single human to move against the main trend. However, in our scenario, persons may move freely. There may be many motion patterns, and a particular individual may follow any one or none of them. The motion patterns are used to explain non-linear trajectories, but do not produce extra penalties for individuals not following them.

Robust appearance models are also key for multi-target tracking, and many methods collect training samples online from tracklets to learn appearance models [2, 5, 6, 8, 15]. However, these approaches collect positive samples from the same tracklets in a few neighboring frames; therefore, the positive samples often lack diversity. Once there are illumination or pose changes, tracklets belonging to the same target but with long gaps may have different appearance models, as positive samples are always from neighboring frames. On the contrary, our method utilizes entry/exit points to find potential correct associations, and positive samples may come from tracklets with long gaps and thus have high diversities. To the best of our knowledge, there has been no explicit use of entry/exit points to improve appearance models in multi-target tracking tasks.

Human interactions have also been considered in tracking. Some focus on using social behaviors to avoid collisions for a few persons [14, 11], or finding moving groups of people to estimate the motion of each individual [10, 19]. However, in crowded scenes, heavy occlusions may cause very close persons to merge. In addition, previous moving groups are used as motion constraints when each member is clearly visible [10, 19]. However we use moving groups to complete tracklets when they are not observable due to the failure of the detector.
3. Motion map learning for motion affinities

In this section, we introduce an online learning approach to find reasonable non-linear motion patterns for each sliding window, and use them to produce more precise motion affinities, i.e., $A_m(\cdot)$ in Eq. 1.

3.1. Non-linear motion map learning

As modern detectors are quite robust, we safely assume that there are no long time continuous missing detections for a target if there are no occlusions. For unoccluded targets, it is easy to associate tracklets based on linear motion assumptions as gaps are often quite small. However, in a long time span, these targets may follow non-linear patterns, which may provide guidance for associations of other tracklets. Similar to [6], we also adopt a multi-level association approach, i.e., gradually associate tracklets in multiple steps instead of one-time association. Non-linear motion patterns learned from tracklets in previous level are used for current level association.

For current sliding window, a motion map $M = \{T_1, T_2, \ldots, T_m\}$ is defined as a set of tracklets that include confident non-linear motion patterns. A tracklet $T_i = \{d_{i1}^t, \ldots, d_{in}^t\}$ is a set of detection responses, or interpolated responses, in consecutive frames, where $t_i^s$ and $t_i^e$ denote the starting and ending frame numbers, and $d_i^t = \{(p_i^t, s_i^t, v_i^t)\}$ denote the response at frame $t$, including position $p_i^t$, size $s_i^t$, and velocity vector $v_i^t$.

Due to possible false alarms or false associations, we build the motion map only on confident tracklets, each of which satisfies two constraints: 1) it is long enough, e.g., longer than 50 frames, as false tracklets are mostly short ones; 2) it is not or only lightly occluded by other tracklets, e.g., at most 10% frames having visibility ratios less than 70%, as most association errors happen when there are heavy occlusions. Otherwise, a tracklet is classified as an unconfident one.

For each confident tracklet, we remove the linear motion parts in the head or in the tail. If the remaining parts still satisfy a non-linear motion pattern, we put it into the motion map. The motion map learning algorithm is shown in Algorithm 1, where $(a, b)$ denotes the angle between vector $a$ and vector $b$, and $(x, y)$ denotes a vector from position $x$ to $y$. The threshold $\theta$ is set to 10 degree in our experiments.

As shown in Figure 2, the learned motion map $M$ is a union of existing non-linear moving tracklets, which are used for explaining non-linear gaps between other tracklets in the next level of association.

3.2. Estimation of motion affinities

In most previous work [18, 3], motion affinity is estimated by a linear motion assumption as shown in Figure 3. The affinity score is given as

$$G(p_{\text{tail}} + v_{\text{tail}} \Delta t - p_{\text{head}}, \Sigma_p)G(p_{\text{head}} - v_{\text{head}} \Delta t - p_{\text{tail}}, \Sigma_p)$$

(2)

where $\Delta t$ is the frame difference between $p_{\text{tail}}$ and $p_{\text{head}}$, and $G(\cdot, \Sigma)$ is the zero-mean Gaussian function.

![Figure 3. Estimation of motion affinity using linear assumptions.](image)

![Figure 4. Estimations of motion affinity using the motion map.](image)

Algorithm 1: The algorithm for learning the motion map.

Input: tracklets from previous level \{T_1, T_2, \ldots, T_n\}

Initialize motion map: \(M = \phi\)

For \(i = 1, \ldots, n\) do:

- If \(T_i\) is not a confident track, continue for next iteration.

  initialize non-linear motion start frame \(t_s = t_e^i\) and end frame \(t_e = t_e^i\).

  For \(j = t_s + 1, \ldots, t_e\) do:

    - if \((p_j'_{\text{tail}}, (p_j'_{\text{head}}, p_j'_{\text{tail}})) > \theta\), then \(t_s = j - 1\), break.

  For \(j = t_s - 1, \ldots, t_e\) do:

    - if \((p_j'_{\text{head}}, (p_j'_{\text{tail}}, p_j'_{\text{head}})) > \theta\), then \(t_e = j + 1\), break.

  if \(t_s < t_e \&\& (p_j'_{\text{tail}}, (p_j'_{\text{head}}, p_j'_{\text{tail}})) > 20 \&\& (p_{\text{head}} - p_j'_{\text{head}}, (p_{\text{tail}} - p_j'_{\text{tail}})) > 20\),

  \(M = M \cup \{T_i\}\), where \(T_i^* = \{d_{i1}^*, \ldots, d_{in}^*\}\).

Output: the motion map \(M\).
shown in Figure 4. If multiple patterns exist, the one that produces highest affinity score is used.

Note that we do not use motion patterns as priors to penalize tracklets not following the patterns like previous work [17, 13, 22]. Targets are not necessarily assumed to follow any non-linear patterns, but once an individual does, we reduce the penalty for that non-linear motion.

Figure 5 shows a non-linear motion example in a real case. From the 2D map, we see that there is a direction change between $T_{18}$ and $T_1$ though they are the same person. The linear motion assumption would produce a very low score for associating these two. However, our approach indicates that a confident tracklet $T_{16}$ well explains the direction change from $T_{18}$ to $T_1$, and therefore gives a high motion affinity between $T_{18}$ and $T_1$.

4. MIL using the entry/exit map

Appearance models play important roles in tracking. Most previous online learning approaches [2, 5, 6, 8] collect positive training samples, i.e., responses belonging to the same target, from the same tracklet within a few frames. However, these responses are likely quite similar and lack diversity. We further collect potential positive pairs from responses in different tracklets with longer gaps according to the estimated entry/exit map, so that the diversity is higher; then a multiple instance learning approach is proposed to get robust appearance models.

4.1. Incremental learning of the entry/exit map

An entry/exit map is a binary image $I$ with 0 denoting entry/exit points and 1 denoting others. The entry/exit points are positions where a target enters or exits in video frames. We do not constraint order; an entry point is also an exit point and vice versa. We limit our approach to static cameras, so that the entry/exit map does not change with time and can be learned incrementally.

We continuously add starting or ending positions of confident tracklets into $E$, and the neighboring regions of these positions are treated as entry/exit points. We assume that all entry/exit points form a convex hull, i.e., all possible points are at borders of a real 3D scene (not necessarily borders on 2D frames), and a target cannot enter or exit in the middle of a scene. Based on this assumption, we continuously update the entry/exit points by removing those inside the convex hull and adding those outside it. Figure 6 shows an example of the update process. The incremental learning algorithm of the entry/exit map is shown in Algorithm 2.

4.2. Learning for appearance models

With the learned entry/exit map, we identify each tracklet as an entry tracklet, an exit one, both, or neither.

**Definition 1** An entry tracklet $T$ starts at any entry/exit point or at the beginning of current sliding window; otherwise, it is a non-entry tracklet.

The definition of an exit tracklet is similar. For tracklets that are both entry and exit ones, we call them completed tracklets; otherwise, we call them uncompleted tracklets.

Ideally, any real track should be a completed tracklet. However, due to false alarms, appearance changes and occlusions, there are usually many uncompleted tracklets. For an uncompleted but confident tracklet, e.g., a non-exit tracklet, it probably needs to be associated with other tracklets so that the linked tracklet would be completed.

Figure 7 gives an illustration for our potential training pairs collection. For a non-exit confident tracklet $T_1$, we collect all tracklets that have motion affinities with $T_1$ higher than a threshold $\gamma$, set to 0.2 in our experiments, as potential correct associations, e.g., $T_3$, $T_4$, and $T_5$. A positive bag $S$ is formed by $S = \{(d_1^1, d_1^2), (d_1^3, d_1^4)\}$, where $d_1^1 \in T_1$, $d_1^2 \in T_3$, $d_1^3 \in T_4$, and $d_1^4 \in T_5$. At least one pair in a positive bag is a correct pair, so that the bags are suitable for MIL. The positive bags provide more positive pairs from long gap tracklets, and make $T_1$ be more probably associated with one of $T_3$, $T_4$, and $T_5$. Note that for the unconfident tracklet $T_3$ ($T_5$ is similar), it is not necessary to associate with one of $T_3$, $T_4$, or $T_5$, as unconfident tracklets may be false alarms, but it may appear in the positive bags of other confident tracklets, e.g., $T_3$. 

Figure 6. Estimation of entry/exit points. Each circle indicates an estimated entry/exit point; red polygons indicate the convex hulls of all estimated points; blue circles indicate later removed points; yellow circles indicate new added ones.

**Algorithm 2** Learning algorithm for the entry/exit map.

**Input**: confident tracklets obtained from previous association level in current sliding window $\{T_1, T_2, \ldots, T_m\}$

If this is the first sliding window, initialize entry/exit point set $E = \emptyset$, and its convex hull $H = \emptyset$

**For** $i = 1, \ldots, m$ **do**:

- If $p_i^1$ is outside $H$, $E = E \cup \{d_i\}$; if $p_i^1$ is outside $H$, $E = E \cup \{d_i\}$.
- Update convex hull $H$ using current $E$.
- **For** $d_i \in E$, if $p_i$ is inside $H$, $E = E - \{d_i\}$.

**Output**: the binary entry/exit map $I$, where $I(p) = 0$ if $\exists p_r \in E$ so that $|p - p_r| < s$, and $I(p) = 1$ otherwise.
For negative training samples, we adopt the approach used in [6]: responses from tracklets having temporal overlaps are used for forming negative training samples. The feature pool as in [6] is used; it is based on color, shape, and texture, and features are extracted from pre-defined regions of human responses, e.g., the color histogram of the upper body.

We adopt the multiple instance learning framework used in [9], and the algorithm is shown in Algorithm 3, where \( x_i \) denotes the number of elements in \( x_i \) and \( L \) is the log-likelihood of bags to be maximized as

\[
L(H) = \sum_i (y_i \log p_i + (1 - y_i) \log(1 - p_i))
\]

The learned classifier \( H \) is used as the appearance model, and the appearance affinity scores are computed as in [6].

**Algorithm 3** Learning algorithm for appearance models.

Input: training bags \( B = \{(x_i = \{x_i^1, x_i^2, \ldots, x_i^L\}, y_i)\} \), where \( x_i^l = \{d_{il}, d_{yl}\} \) indicates a pair of responses, and \( y_i \in \{1, 0\} \). Feature pool \( F = \{h_1, h_2, \ldots, h_K\} \). Let the number of selected features be \( T (T < K) \).

Initialize classifier function \( H = 0 \)

For \( i = 1, \ldots, T \) do:

- For \( k = 1, \ldots, K \) do:
  - \( p_{ik}^l = 1/(1 + \exp(-(H(x_i^l) + h_k(x_i^l)))) \)
  - \( p_{ik}^l = 1 - \prod_{j \neq i} (1 - p_{ik}^j)^{1/|x_i|} \)
  - \( \omega_{ij} = (y_i - p_{ik}^l)p_{ik}^l/\{x_i^l p_{ik}^l\} \)
- Find \( k^* = \arg\max_k \sum_j \omega_{ij} h_k(x_j^l) \); \( h_t = h_{k^*} \)
- Find \( \alpha_t = \arg\max_k L(H + \alpha h_t) \) by linear search; \( \alpha_t = \alpha^* \)
- \( H = H + \alpha_t h_t \)

Output: \( H = \sum_{t=1}^T \alpha_t h_t \)

5. Track completion

After tracklet associations, there are likely several uncompleted but confident tracklets. These tracklets are probably due to occlusions by targets in other tracklets or in cluttered background where the detector fails. Therefore, it is difficult to observe them from detection responses. Fortunately, people often move in groups, and the group motion can provide priors for motions of each member in the group.

**Definition 2** A moving group is a group of people who move at similar speeds and in similar directions as well as keep close to each other.

Two tracklets \( T_i \) and \( T_j \) belong to the same moving group if they satisfy the following constrains (assuming \( T_j \) is equal or longer than \( T_i \)):
[20, 6]: recall & precision, showing detection performance after tracking; false alarms per frame (FAF); mostly tracked (MT) & mostly lost (ML), the ratio of tracks with successfully tracked parts for more than 80% or less than 20% respectively; partially tracked (PT), 1-MT-ML; fragments (Frag), the number of times that a ground truth trajectory is interrupted; id switches (IDS), the number of times that a tracked trajectory changes its matched id. All data used in our experiments are publicly available2.

6.1. Entry/exit map estimation

The estimations of entry/exit maps for all five scenes used in our experiments are shown in Figure 9. We can see that with time, our approach produces more and more precise estimations. Note that the maps are used for improving tracking not for scene understanding; for some entry/exit points, if no targets appear or disappear there, they would have no influence on tracking. Therefore, the precision is more important than recall. We can see that nearly all our estimated points are correct positions. Some imprecise estimations occur in top parts of images in the second row, because humans are really small on top parts of this scene and detectors would fail there. Therefore, the tracklets often end before reaching the top parts. However, as there are almost no detection responses there, such estimation would have little negative influence on the tracking performance.

6.2. Performance on less crowded data sets

PETS 2009 and CAVIAR are two commonly used data sets for multi-target tracking. Scenes are not crowded, and people are sometimes occluded by other humans or other objects. The PETS 2009 data are the same as used in [1], but we modify the ground truth annotations, so that people who are fully occluded for many frames but are visible later are labeled with the same id. For CAVIAR, we test our approach on the same 20 video clips used in [5, 6].

The comparison results are shown in Table 2. We can see that our approach produces obvious improvements; fragments are greatly reduced on both data sets by over 50% and 35% respectively, while keeping other scores competitive or with some improvements. Some visual results are shown in Figure 10(a). We can see that for almost totally overlapped persons, our tracker does not confuse their identities and finds the correct associations.

6.3. Performance on Trecvid 2008 data set

As the PETS 2009 and CAVIAR data sets are relatively easy, we show more results on the difficult Trecvid 2008 data set. There are frequently heavy occlusions, many non-linear motions, and interactions between people. There are three different scenes in the test videos with three 5000 frames video clips for each scene.

The quantitative comparison results are shown in Table 3. To show effectiveness of each component of our approach, except for final performance, we also report three additional results where only one component of our approach is activated in each. In addition, to see the effectiveness of our estimation of entry/exit maps, we also report the performance using manually assigned maps.

Figure 10(b)-(d) show visual results of our approach. In Figure 10(b), a woman has non-linear motions and is almost fully occluded by person 170 around frame 4262. Tradi-

2http://iris.usc.edu/people/yangbo/downloads.html
tional linear motion assumptions cannot connect tracklets before and after the occlusion; however, with the help of an online learned motion map, we successfully associate these tracklets into one. In Figure 10(c), person 41 is fragmented from frame 2075 to frame 2105 because of heavy occlusions, and his appearance changes from almost black to dark blue due to illumination change. However, our MIL algorithm is able to produce high appearance similarity between the two tracklets, and links them successfully. In Figure 10(d), our approach finds that person 50, 51, and 54 form a moving group. After frame 1265, only person 51 is visible; however, we complete trajectories of person 50 and 54 according to the visible tracklet 51.

### 6.4. Computation speed

The speed is highly related with the number of targets in a video. Our approach is implemented using C++ on a PC with 3.0GHz CPU and 8GB memory. The average speeds are 48 fps, 16 fps, and 6 fps for CAVIAR, PETS 2009, and Trecvid 2008, respectively. Comparing 48 fps and 7 fps for CAVIAR and Trecvid 2008 reported in [6], our approach does not provide much extra computational burden. In our experiments, most of the computation is spent on feature extraction, followed by low-level association.

### 7. Conclusion

We described an online learning approach for multi-target tracking. We learn a non-linear motion map, describe a multiple instance learning algorithm for better appearance models based on estimated entry/exit points, as well as complete tracklets based on moving groups. Our approach improves the performance a lot compared with up-to-date approaches, while adding little extra computation cost.

### Acknowledgment

This paper is based upon work supported in part by Office of Naval Research under grant number N00014-10-1-0517.

### References


[22] X. Zhao and G. Medioni. Robust unsupervised motion pattern inference from video and applications. In ICCV, 2011. 2, 4

---

3Detection time is not included in both our speed and that in [6]
Table 3. Comparison of results on TRECVID 2008 dataset. The human detection results are the same as used in [20, 5, 6], and are provided by courtesy of authors of [6].

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>FAF</th>
<th>GT</th>
<th>MT</th>
<th>PT</th>
<th>ML</th>
<th>Frag</th>
<th>IDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF Tracking [20]</td>
<td>79.2%</td>
<td>85.8%</td>
<td>0.996</td>
<td>919</td>
<td>78.2%</td>
<td>16.9%</td>
<td>4.9%</td>
<td>319</td>
<td>253</td>
</tr>
<tr>
<td>OLDAMs [5]</td>
<td>80.4%</td>
<td>86.1%</td>
<td>0.992</td>
<td>919</td>
<td>76.1%</td>
<td>19.3%</td>
<td>4.6%</td>
<td>322</td>
<td>224</td>
</tr>
<tr>
<td>PRIMPT [6]</td>
<td>79.2%</td>
<td>86.8%</td>
<td>0.920</td>
<td>919</td>
<td>70.0%</td>
<td>17.7%</td>
<td>5.2%</td>
<td>283</td>
<td>171</td>
</tr>
<tr>
<td>Motion map only</td>
<td>79.5%</td>
<td>87.8%</td>
<td>0.853</td>
<td>919</td>
<td>75.1%</td>
<td>19.2%</td>
<td>3.7%</td>
<td>235</td>
<td>155</td>
</tr>
<tr>
<td>MIL only</td>
<td>80.2%</td>
<td>87.8%</td>
<td>0.855</td>
<td>919</td>
<td>77.1%</td>
<td>17.7%</td>
<td>5.2%</td>
<td>250</td>
<td>167</td>
</tr>
<tr>
<td>Track completion only</td>
<td>80.1%</td>
<td>87.3%</td>
<td>0.895</td>
<td>919</td>
<td>76.7%</td>
<td>17.6%</td>
<td>5.7%</td>
<td>258</td>
<td>169</td>
</tr>
<tr>
<td>Our approach</td>
<td>80.2%</td>
<td>75.5%</td>
<td>0.880</td>
<td>919</td>
<td>76.9%</td>
<td>17.6%</td>
<td>5.5%</td>
<td>242</td>
<td>153</td>
</tr>
<tr>
<td>Our approach w/ manual maps</td>
<td>79.5%</td>
<td>87.1%</td>
<td>0.903</td>
<td>919</td>
<td>77.0%</td>
<td>17.5%</td>
<td>5.5%</td>
<td>235</td>
<td>152</td>
</tr>
</tbody>
</table>

Figure 10. Examples of tracking results of our approach on PETS 2009, CAVIAR and Trecvid 2008 data sets.