Unsupervised Trajectory Modeling based on Discrete Descriptors for Classifying Moving Objects in Video Sequences

Damián Campo
Mohamad Baydoun
Lucio Marcenaro
Andrea Cavallaro
Carlo Regazzoni

October 2018
1. Introduction to the main problem
2. Proposed method
   2.1. Offline vocabulary learning
   2.2. Incremental classification process
3. Employed datasets and experimental results
4. Conclusions and future work
Introduction to the problem

- Is it possible to have a set of descriptors that encode observed motions?

- Is it possible to distinguish trajectories with different dynamics appearing in the same location?

- Is it possible to classify the observed trajectories incrementally, i.e., as observations arrive?
Proposed method

Video frame in timestamp $k$

For a single object

Location in the scene (2-dimensional)

Velocity information (2-dimensional)

Time spent in the video sequence (1-dimensional)

$$X_k = \begin{bmatrix} x_k & y_k & \dot{x}_k & \dot{y}_k & t_k \end{bmatrix}$$
Clustering of similar state information

Video sequence

State sequences of objects

\[
\begin{align*}
X_k^1 & \quad X_{k+1}^1 \\
X_k^2 & \quad X_{k+1}^2 \\
\vdots & \quad \vdots \\
X_k^N & \quad X_{k+1}^N
\end{align*}
\]

Identified objects

\[
\begin{align*}
X_{K_1}^1 & \quad X_{K_1}^1 \\
X_{K_2}^2 & \quad X_{K_2}^2 \\
\vdots & \quad \vdots \\
X_{K_N}^N & \quad X_{K_N}^N
\end{align*}
\]

Trajectories

State sequences of objects

State sequences of objects

Self organizing map (SOM) training

\[
\omega_{SOM} = [\beta, \alpha, \gamma]
\]

Clusters of states (vocabulary)

\[
C = \{C_1, C_2, \ldots, C_M\}
\]

5-dimensional regions encoding objects’ dynamics

\[
\beta + \alpha + \gamma = 1
\]

\(\beta\): Location weight

\(\alpha\): Velocity weight

\(\gamma\): Spent time weight
Vocabulary properties

**Vocabulary**

\[ C = \{C_1, C_2, \ldots, C_M\} \]

5-dimensional regions encoding objects’ dynamics

**Letters**

\[
C_m = \begin{bmatrix} \bar{x}_m \\ \bar{y}_m \\ \bar{x}_m \\ \bar{y}_m \\ \bar{t}_m \end{bmatrix}
\]

Where \( C_m \in C \)

**Distance between letters**

\[
d_{i,j} = (\omega_{SOM} A) \ \text{abs}(C_i - C_j).
\]

Where:

\[
\omega_{SOM} = [\beta, \alpha, \gamma] \ ; \ A = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix};
\]

\( C_i \in C \) and \( C_j \in C \)

A distance matrix \( D \) containing the separation between letters is defined as:

\[
D = \begin{bmatrix} 0 & \cdots & d_{i,j} \\ \vdots & \ddots & \vdots \\ d_{i,j} & \cdots & 0 \end{bmatrix}
\]
Words generation

Any 5-dimensional state $X_k$ can be transformed into a vocabulary letter by following the function $G(X_k)$, defined as:

$$G(X_k) = m_{X_k} = \arg \min_m ((\omega_{SOM} A) \ abs(X_k - C_m))$$

Let a trajectory $l$ be defined as:

$$\tilde{X}_l = \{X_1^l, ..., X_{k_l}^l, ..., X_{K_l}^l\}$$

Discrete version of trajectory $l$:

$$m_{\tilde{X}_l} = \{m_{X_1^l}, ..., m_{X_{k_l}^l}, ..., m_{X_{K_l}^l}\}$$

“Word” (class) generation

Given a set of trajectories:

$$\tilde{X}_{tot} = \{\tilde{X}_1, ..., \tilde{X}_l, ..., \tilde{X}_L\}$$

How to generate a dictionary (set of words) that encodes observed dynamics for classification purposes?

Can such dictionary be built as trajectories are observed one by one?

Can such dictionary be quasi-invariant to the order of observed trajectories?
Incremental dictionary creation

Set of trajectories
\[ \widetilde{X}_{tot} = \{\widetilde{X}_1, ..., \widetilde{X}_l, ..., \widetilde{X}_L\} \]

Select a randomly a new trajectory \( \widetilde{X}_l \)

Obtain a set of activated letters: \( m_{\widetilde{X}_1} \)

Update the class \( f(\text{min}) \) by adding letters of \( m_{\widetilde{X}_1} \) that were not in such class

\( \text{score}_{f(\text{min})} \): Minimum class score
\( \theta \): Threshold value

For identified class \( f \):

\( d_{\text{min}} \): Minimum distances between letters of class \( f \) and \( m_{\widetilde{X}_1} \) based on matrix \( D \)

\[ \text{score}_f = \frac{\text{sum}(d_{\text{min}})}{R_f} \]

\( R_f \): Number of letters in class \( f \)

Creation of new class defined as \( m_{\widetilde{X}_1} \)
Summarizing (vocabulary creation)

Offline vocabulary learning

Video sequences $\rightarrow$ Generation of entities' trajectories $\rightarrow$ Self-organizing map $\rightarrow$ Vocabulary & Distance matrix

$\bar{X}_{tot} = \left\{ \bar{X}_1, \ldots, \bar{X}_L \right\}$

$D = \begin{bmatrix} 0 & \cdots & d_{i,j} \\ \vdots & \ddots & \vdots \\ d_{i,j} & \cdots & 0 \end{bmatrix}$
Summarizing (dictionary creation)

**Incremental classification process**

- Entity’s trajectory
  - Trajectory creation
  - Analysis of activated vocabulary letters
  - $\theta, D$
  - Creation of new class
  - Yes
  - No
  - Update existing class model
  - Generate new class model
  - (Trajectory classes)
Simulated data

The CROSS dataset [1] is a simulated environment where objects move according to 19 classes (words) proposed by authors.

- Each class contains 100 tracks designed for training models and 500 trajectories for testing them.
- Training tracks are used to build the vocabulary.
- Testing trajectories are used to generate the dictionary (classes).

Results

The proposed algorithm found 47 trajectory classes (words) in an unsupervised way. Such number differs from the 19 proposed classes due to the inclusion of velocity and time spent in the video sequences.

Subclasses generated for three ground truth classes

Subclasses generated for six ground truth classes
Confusion matrix for simulated data

Maximum confusion is obtained between couples of classes 3-7 and 11-13
Maximum confusion cases

Random testing trajectories
classes **7-3**

Random testing trajectories
classes **11-13**
The PDTV dataset [2] consists of video sequences containing 51 trajectories organized in 16 classes (words) proposed by authors.

All 51 trajectories are used to build the vocabulary and the dictionary.


Results

As proposed by the ground truth, 16 classes (words) were also obtained by our algorithm. Nonetheless, the acquired labels are slightly different from the ground truth.
Comparison with proposed labels

Ground truth labels

Obtained labels

Identification of high velocity class

Grouping similar car motions

Grouping similar pedestrian motions
Advantages of proposed method

Vehicles stopped by the traffic light for long periods of time represent a problem for other classification algorithms [3] since they share many similar points.

Conclusions and future work

Our method for unsupervised trajectory clustering that uses a weighted SOM to generate a common vocabulary that encodes relevant trajectory information.

A distance matrix from the produced vocabulary to facilitate the incremental recognition of trajectory patterns (words) that can be used for classifying unobserved trajectory data. Results obtained with real and simulated data suggest that our method can generate detailed trajectory classes automatically.

Our approach enables the obtainment of a dictionary of trajectories based on their location, velocities and time spent a video sequence.

As a future work, we will employ probabilistic filtering that uses continuous and discrete information for tracking of objects in video data.
Thank you for your attention