



Orange Labs - Rennes, France
Multimedia Contents Analysis Technologies
www.francetelecom.com/rd/



LIRIS, CNRS / INSA Lyon - Lyon, France
Feature Extraction and Identification
<http://liris.cnrs.fr>

Action Classification in Soccer Videos with Long Short-Term Memory Recurrent Neural Networks

Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia and Atilla Baskurt

September 18th 2010

- Automatic extraction of high-level (semantic) information from videos.
- Action recognition is a crucial task to semantically describe a video sequence.
- Most existing methods in action recognition make no use of the temporal / motion information of the video sequence :

	Motion information	Temporal information
Key frame	✗	✗
2D features	✗	✓
3D features	✓	✗

- Almost all existing methods use SVM classifiers despite they were not primarily designed for sequence classification.
- Sports videos are particularly interesting contents due to their high commercial potential.

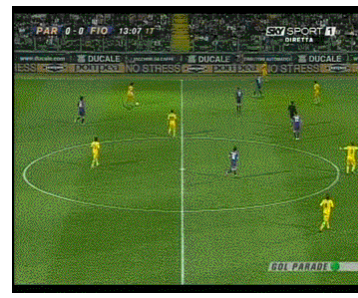
- Provide a classification scheme which takes into account visual, temporal and motion information.
- Experimentally compare RNN and SVM classifiers for this application (previous work by Ballan et al. [*Ballan'09*]).
- All the experiments will be carried out on the MICC-Soccer-Actions-4 dataset



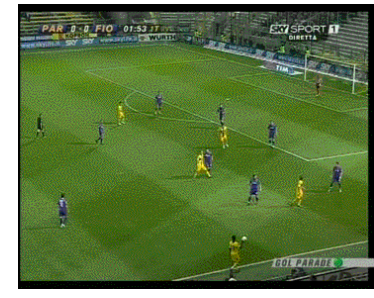
Goal kick



Placed kick

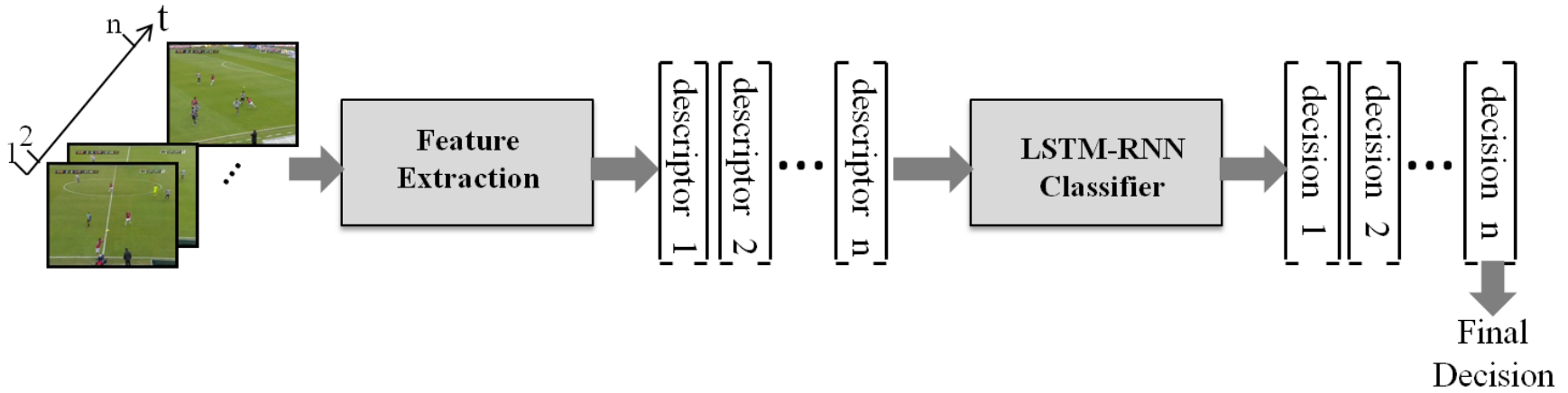


Shot on goal



Throw-in

- Video lengths between 100 and 2500 frames → use a modified RNN scheme which can handle long sequences.



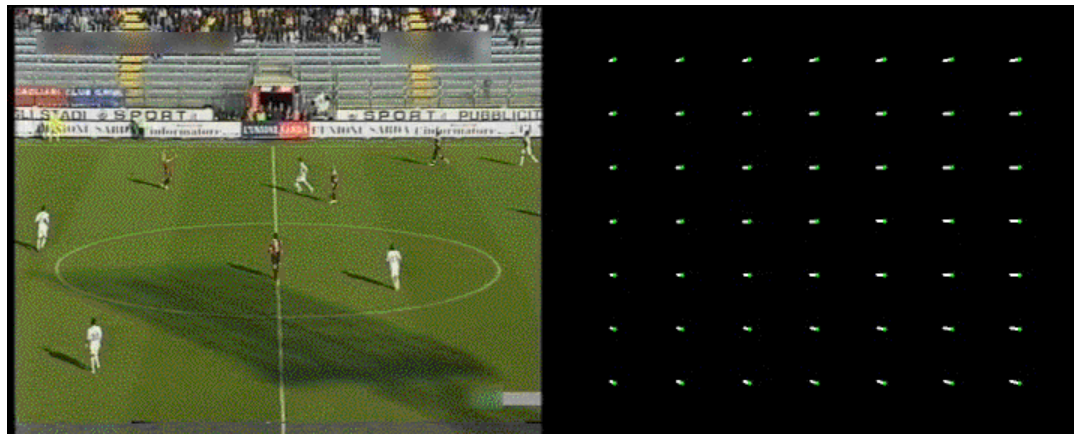
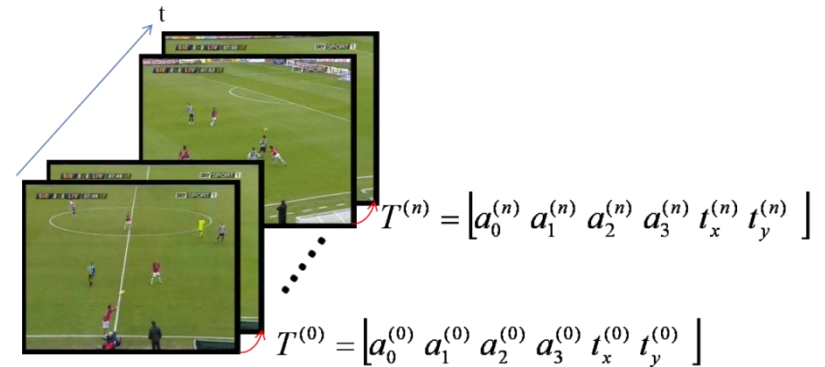
- A video = a sequence of descriptors (one descriptor per image) corresponding to a set of features.
- An LSTM-based Recurrent Neural Network is trained to recursively classify the sequences.
- The final decision corresponds to the accumulation of several individual decisions.

- Feature extraction for action classification
- Action classification using LSTM-RNN
- Experimental results

- Feature extraction for action classification
- Action classification using LSTM-RNN
- Experimental results

- Main idea :
 - Each image of the video is viewed as a frequencies histogram of visual words.
 - We used SIFT salient points to generate a codebook of 30 visual words (empirical choice).
 - The video is represented by a sequence of histograms which allows :
 - To encode the visual content of the scene (visual information).
 - To model transitions between images (temporal information).
- Ballan et al. [*Ballan'09*] used the same representation with a string kernel SVM classifier → Performance comparison with our neural classification scheme.
- Output : a visual descriptor vector of size 30 per image.

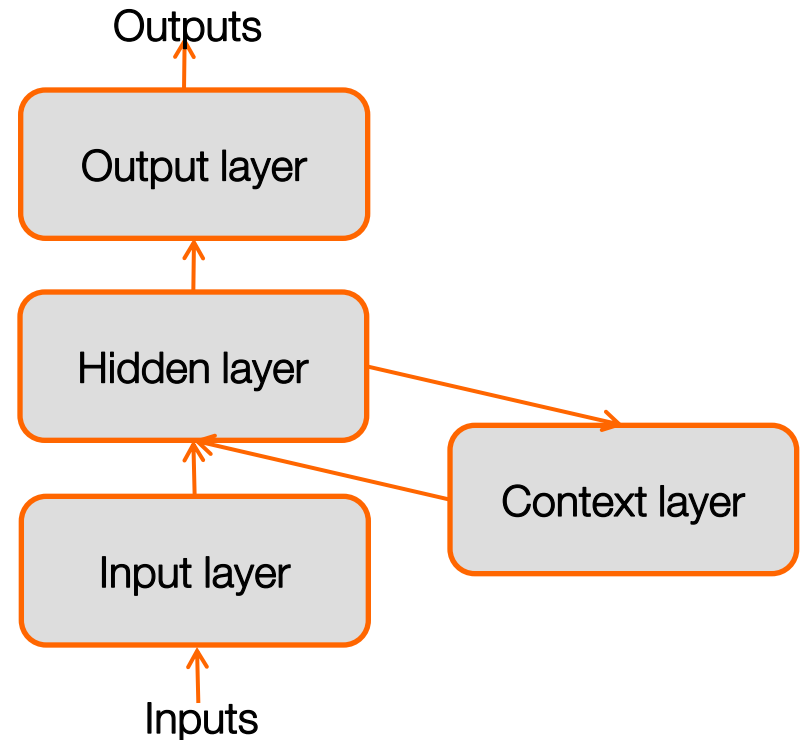
- Goal : Estimate the motion represented by the largest number of elements of the scene : the camera motion for soccer videos.
- A video = a succession of affine transformations between consecutive images.
- We use SIFT matches to estimate transformation parameters (mean square estimation).
- RANSAC algorithm : Outliers are ignored.

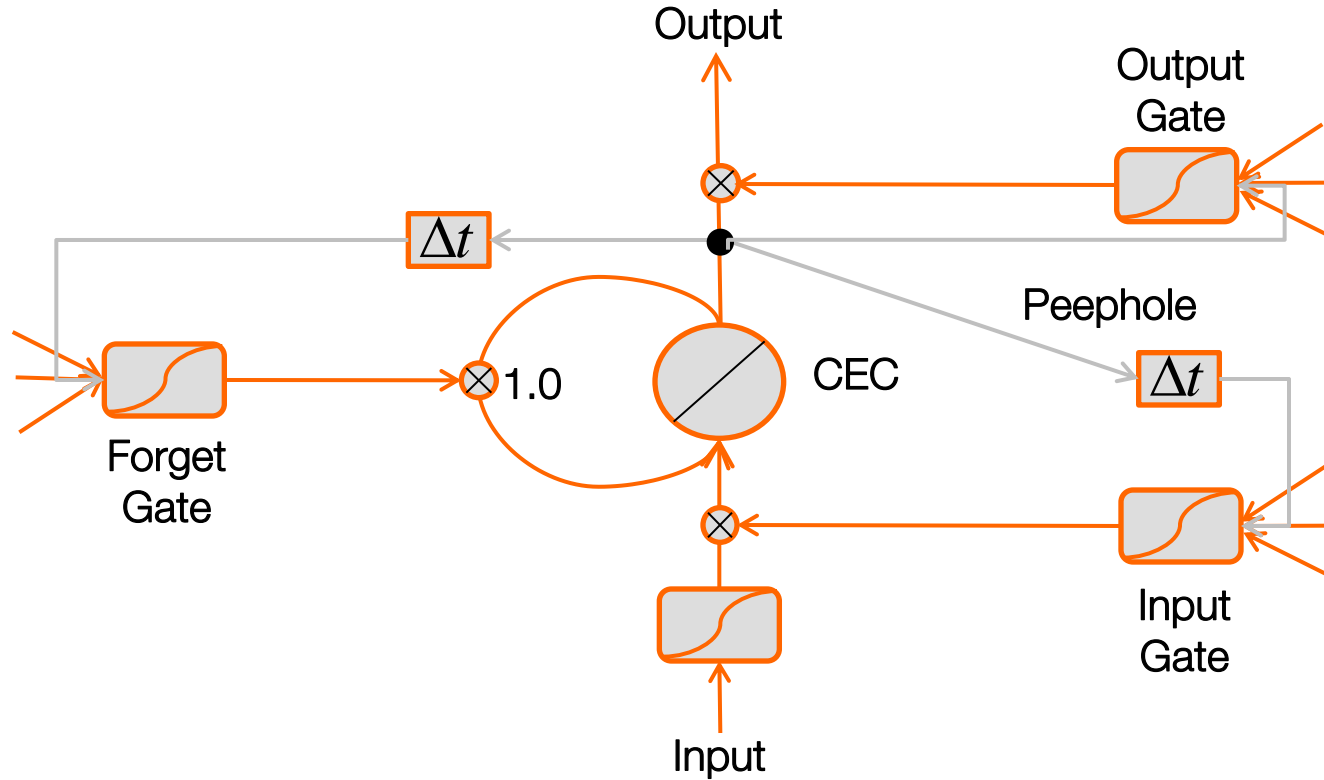


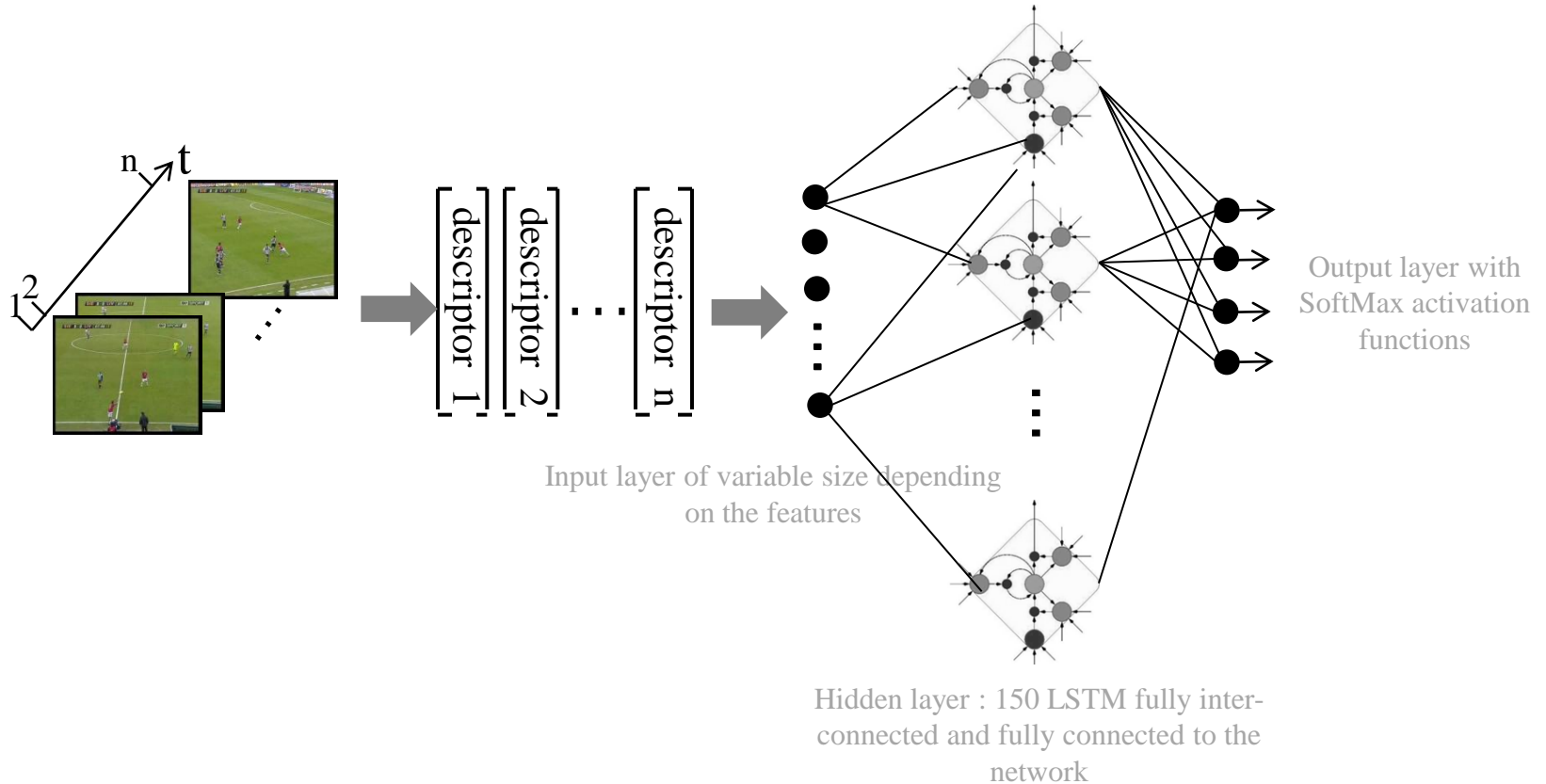
- Output : a motion descriptor vector of size 6 per image.

- Feature extraction for action classification
- Action classification using LSTM-RNN
- Experimental results

- RNN = ANN with a **context layer** which allows to **remember** and **use** previous inputs.
- Classical RNNs are not able to deal with some sequence processing tasks because of their insufficient **short-term memory** [*Schmidhuber'91*].
- **Vanishing gradient problem** : in practice, RNNs fail to learn tasks which involve time lags > 50 timesteps between inputs and corresponding targets.
- Schmidhuber et al. introduced a particular recurrent architecture which was especially designed to overcome this drawback : the **Long Short-Term Memory**.







- About 10^5 trainable weights, depending on the Input size.
- Online-BPTT gives better generalization results than RPROP-based batch learning.

- Feature extraction for action classification
- Action classification using LSTM-RNN
- Experimental results

- 3-fold cross validation for all the experiments.
- Sequences distribution between training / test :

	Training	Test
Config. 1	68 (17/class)	32 (8/class)
Config. 2	68 (17/class)	32 (8/class)
Config. 3	64 (16/class)	36 (9/class)

- We added vertically flipped versions of each video of the training set to increase the examples number.

- LSTM input size is 30 per timestep.
- Evaluation compared to previous work with other classifiers :

	Classification rate
BoW + k-NN <i>[Ballan'09]</i>	52.75%
BoW + string kernel SVM <i>[Ballan'09]</i>	73.25%
BoW + LSTM-RNN	76%

	Goal-kick	Placed-kick	Shot-on-goal	Throw-in
Goal-kick	0,92	0,08	0	0
Placed-kick	0,08	0,8	0	0,12
Shot-on-goal	0	0,2	0,72	0,08
Throw-in	0,12	0,12	0,16	0.6

- The neural classification scheme largely outperforms the k-NN based approach and gives better results than the SVM based one.

- LSTM input size is 6 per timestep.

	Classification rate
Dominant motion + LSTM-RNN	77%

	Goal-kick	Placed-kick	Shot-on-goal	Throw-in
Goal-kick	0.64	0.28	0.08	0
Placed-kick	0.08	0.68	0.08	0.16
Shot-on-goal	0.08	0	0.88	0.04
Throw-in	0.08	0	0.04	0.88

- Motion information is as discriminative as visual one for this type of actions.
- A surprisingly good result : 77% of the actions are recognized relying only on camera motion.

- Results are complementary :
 - BoW : best results for “Placed kick” (92%) and “Goal kick” (80%).
 - Dominant motion : best results for “Shot on goal” (88%) and “Throw in” (88%).
- Concatenation of the two descriptors for each timestep : input size is 36.

	Classification rate
Dominant motion + BoW + LSTM-RNN	92%

	Goal-kick	Placed-kick	Shot-on-goal	Throw-in
Goal-kick	1	0	0	0
Placed-kick	0.04	0.84	0.08	0.04
Shot-on-goal	0	0.12	0.88	0
Throw-in	0.04	0	0	0.96

- LSTM can handle information of different nature.
- 92% is the best published result on this dataset.

- The proposed neural approach is superior, for this application, to k-NN and string kernel SVM-based ones.
- Motion descriptors contain as much discriminant information as visual ones for this type of actions.
- The combination of the two information leads to a classification rate of 92%, which is the best published result on the MICC-Soccer-Actions-4 dataset.
- LSTM-RNN are able to learn to classify variable length sequences, managing features of different nature.

Future work :

- We are currently working on an hybrid approach including an automatic feature extraction step using 3D Convolutional Neural Networks. This would improve the genericity of the approach.
- We plan to test the approach on other more complex sports/scenarios.

Thank you

Moez Baccouche

<http://liris.cnrs.fr/moez.baccouche/>
moez.baccouche@orange-ftgroup.com