Recurrent Space-time Graph Neural Networks

Andrei Nicolicioiu*, Iulia Duta*, Marius Leordeanu

Bitdefender, Romania

summarized by **Seongho Choi**

Seoul National University

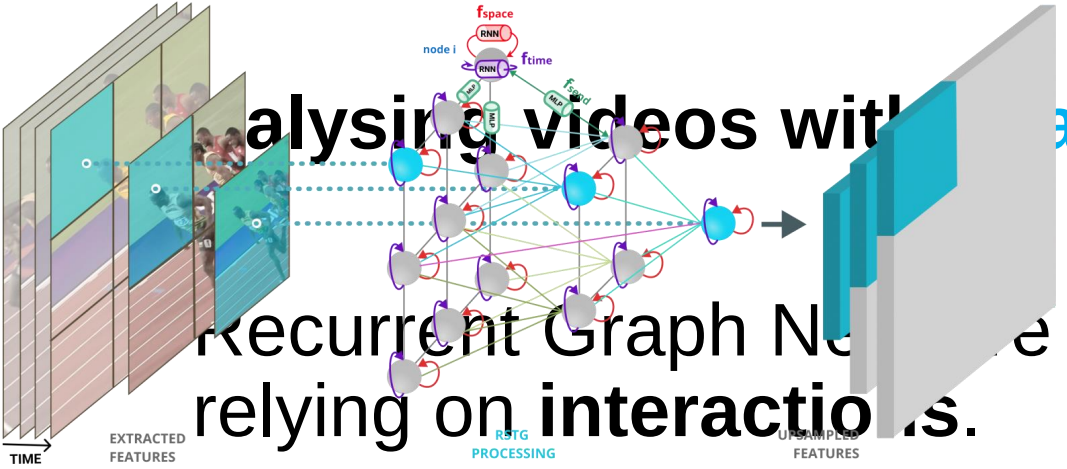
*Equal contribution

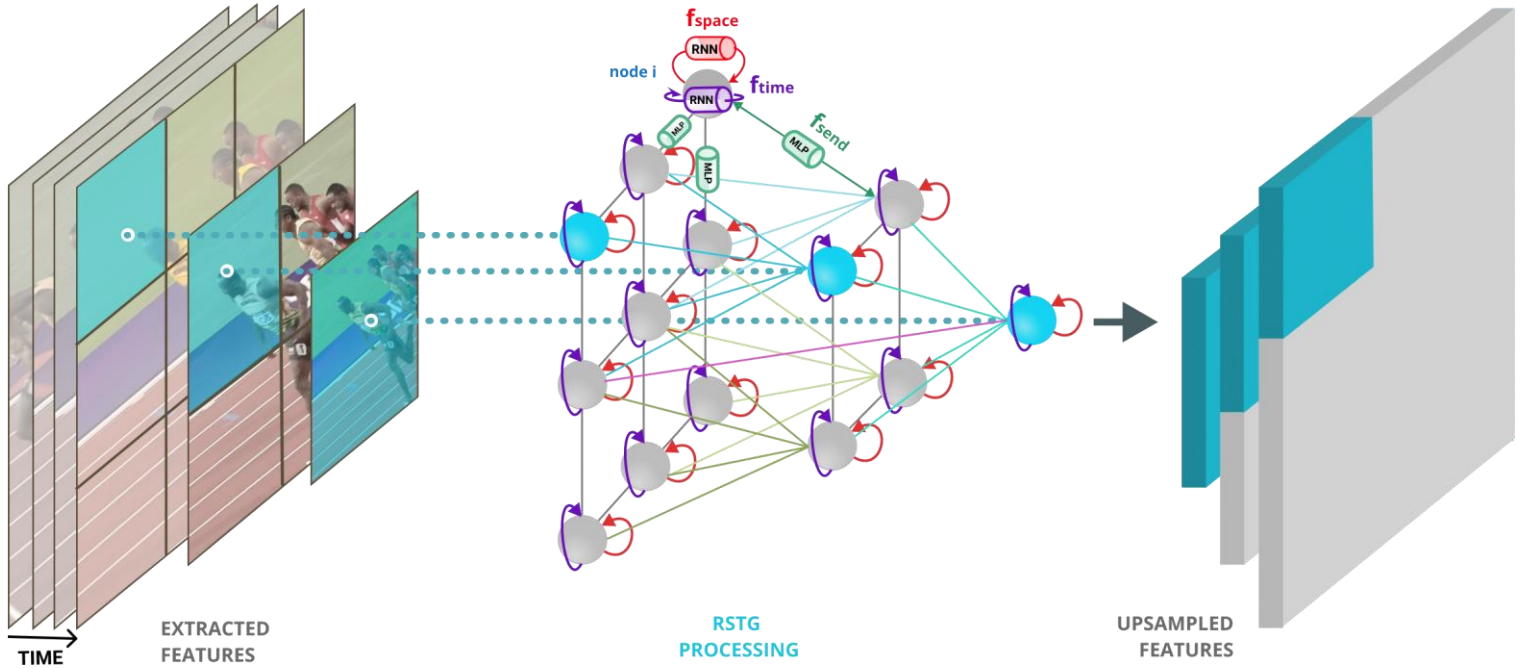
Understanding video

- **spatial** interactions happening at the frame level
- **temporal** interactions between over time
- **long-range** interactions between distant frames

Analysing videos with ...atio-temporal graph models

Recurrent Graph Ne... ...e suited for video analysis tasks heavily relying on **interactions**.





Algorithm 1 Space-time processing in RSTG

Input: Features $F \in R^{T \times H \times W \times C}$

repeat

$\mathbf{v}_i \leftarrow extract\_features(F_t, i)$ $\forall i$

for $k = 0$ **to** $K - 1$ **do**

$\mathbf{v}_i = \mathbf{h}_i^{t,k} = \mathbf{f}_{time}(\mathbf{v}_i, \mathbf{h}_i^{t-1,k})$ $\forall i$

$\mathbf{m}_{j,i} = \mathbf{f}_{send}(\mathbf{v}_j, \mathbf{v}_i)$ $\forall i, \forall j \in N(i)$

$\mathbf{g}_i = \mathbf{f}_{gather}(\mathbf{v}_i, \{\mathbf{m}_{j,i}\}_{j \in N(i)})$ $\forall i$

$\mathbf{v}_i = \mathbf{f}_{space}(\mathbf{v}_i, \mathbf{g}_i)$ $\forall i$

end for

$\mathbf{h}_i^{t,K} = \mathbf{f}_{time}(\mathbf{v}_i, \mathbf{h}_i^{t-1,K})$ $\forall i$

$t = t + 1$

until end-of-video

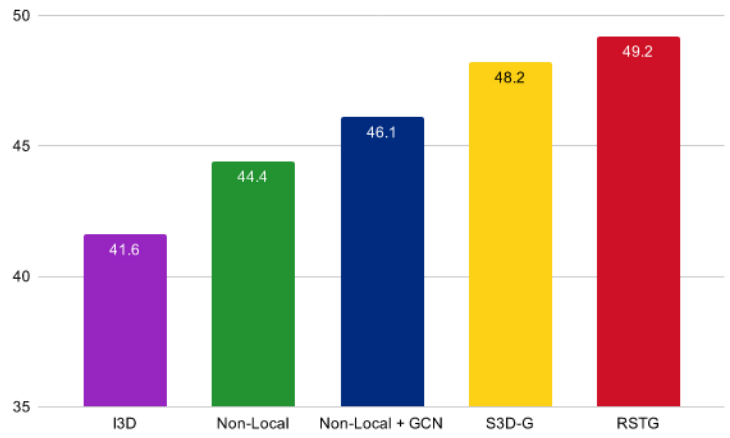
$\mathbf{v}_{final} = f_{aggregate}(\{\mathbf{h}_i^{1:T,K}\}, i)$

Real world experiments

- Something-Something v1
  - human-object interaction dataset
  - interactions between entities across the entire video are essential
- RSTG shows state of the art performance

a. Failing to put smt into smt because smt does not fit

b. Pretend to put smt into smt



Summary & Limitation

- RSTG (Recurrent Space-time Graph Neural Networks)
  - factorize space and time and process them differently
  - achieves a relatively low computational complexity
  - shows state-of-the-art performance on real world dataset

- Limitation
  - not various experiments on other dataset
  - not modeling long-range interactions