Hierarchical Relational Networks for Group Activity Recognition and Retrieval

Mostafa S. Ibrahim and Greg Mori

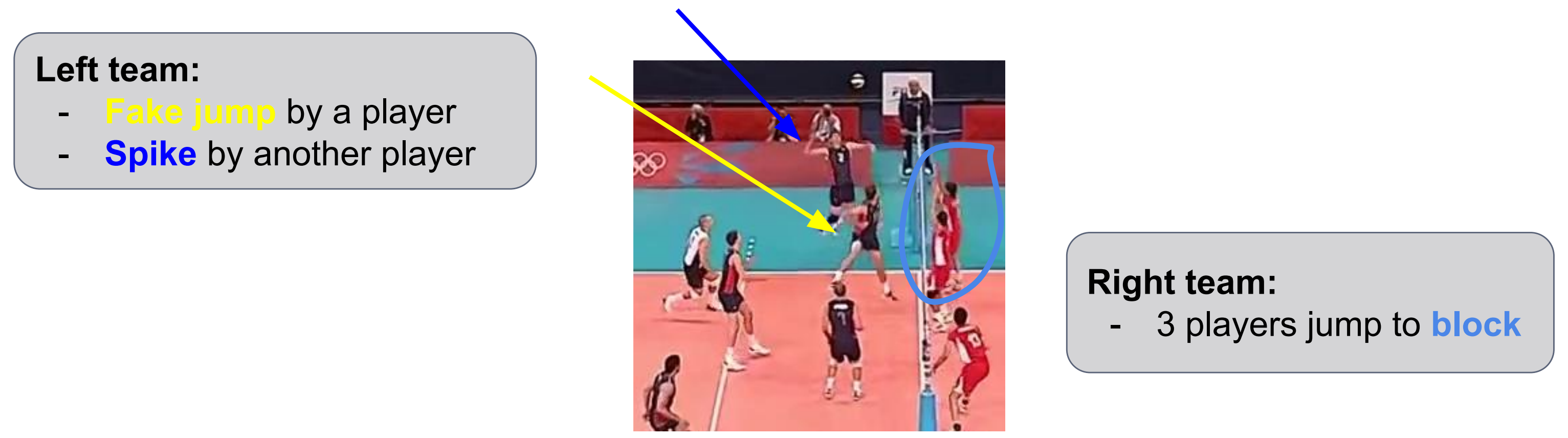
Simon Fraser University

Vision and Media Lab

ECCV 2018 European Conference on Computer Vision
8 – 14 September 2018 | Munich, Germany

Motivation

- Modeling **structured relationships** between people in a scene
- **Generic Network Module** for representing Hierarchical Relationships

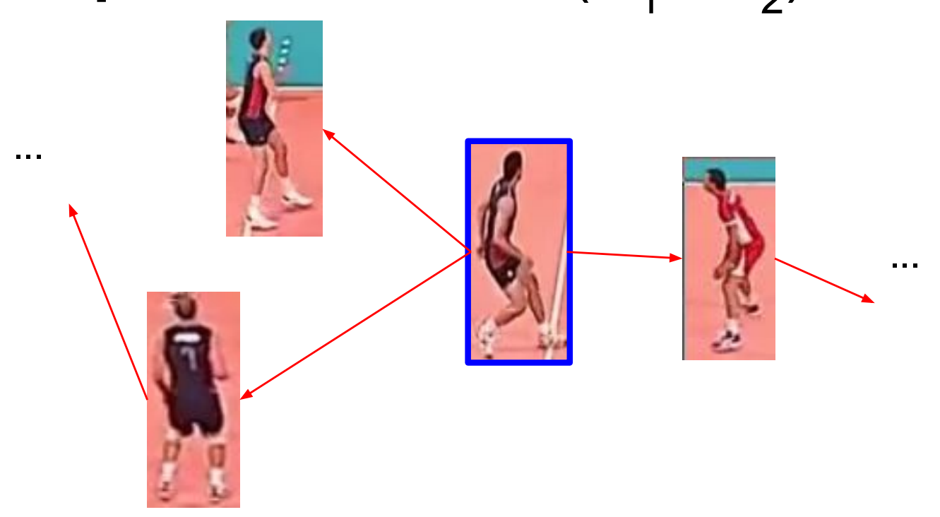


Left team:
- Fake jump by a player
- **Spike** by another player

Right team:
- 3 players jump to **block**

Contribution: Relational Layer

- **Context:** K objects (e.g. players in volleyball) interacting together
- **Input**: K feature vectors (each of size N_1) + **potential graph** of their relationships
- **Output**: K feature vectors (each of size N_2): relationship-based object representation
- Layer **parameters** is $O(N_1 \cdot N_2)$, independent from number of objects



Relationships given as a **graph**

Contribution: Hierarchical Relational Network

- **Stack** multiple relational **layers**, each layer is associated with a relationship graph
  - Output of a relational layer is fed to the next one
- **Network Output**: K feature vectors encoding *hierarchical relational information*
- **Relational Autoencoder** model for compact scene representation
  - **Denoising** Autoencoder variant to infer missing objects

Applications

- **Supervised Learning: Group Activity Recognition:** People in scene are doing collective activity (e.g. right team is blocking, left team got a win-point)
- **Unsupervised Learning: Action and Scene Retrieval**: Given a scene of (possibly missed) actions, find a scene of similar overall actions.

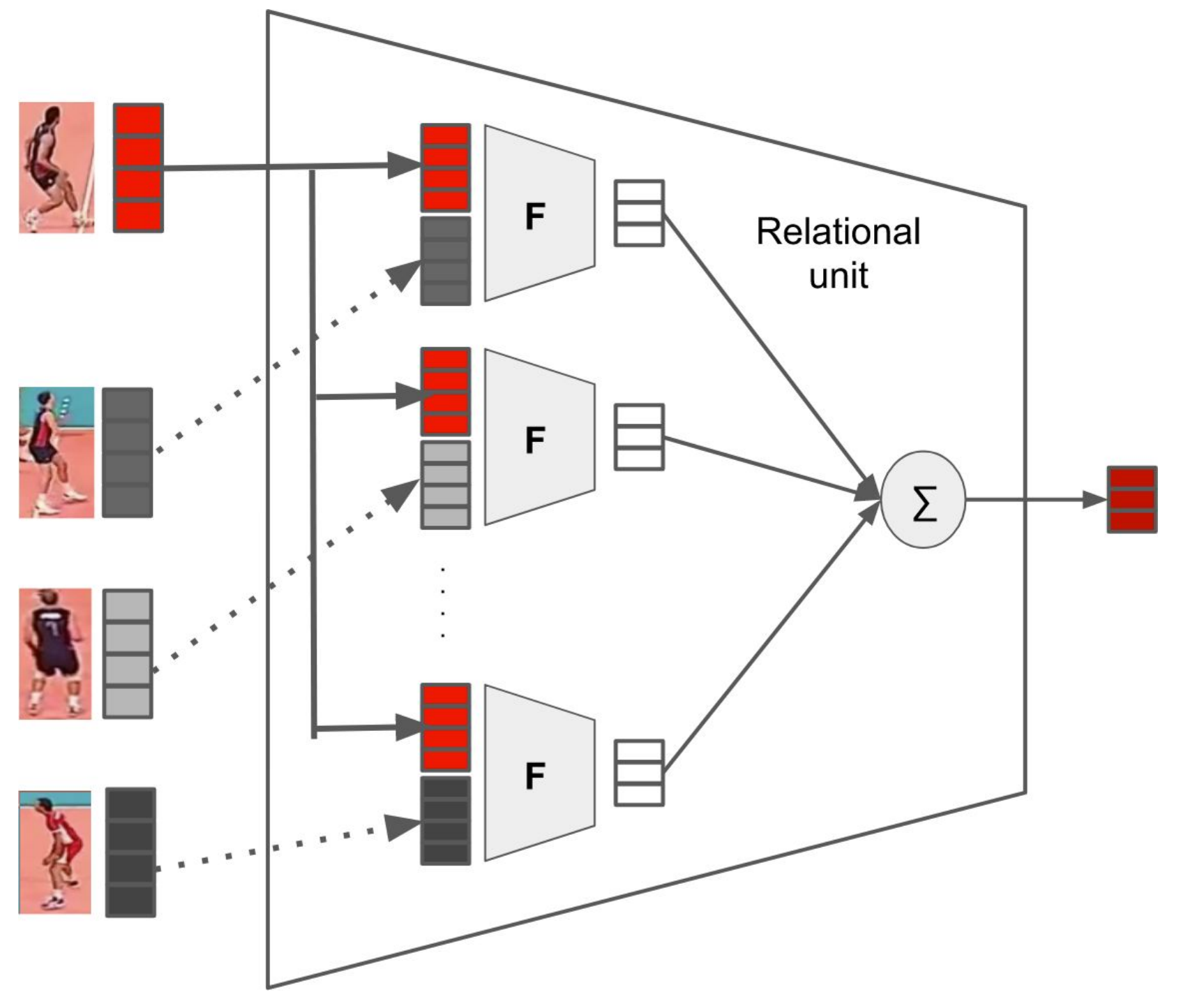
Initial Person Representation

- **Track** a bounding box of each person for a fixed temporal window
- Extract fc7 representation from **VGG19** network for each bounding box per timestep

Relational Unit per layer

- Learn **shared MLP**: receives 2 persons representation and **infers** their edge representation
- **Person's relational representation**: infer all edges representations of a person with his neighbours, then **sum pool** them

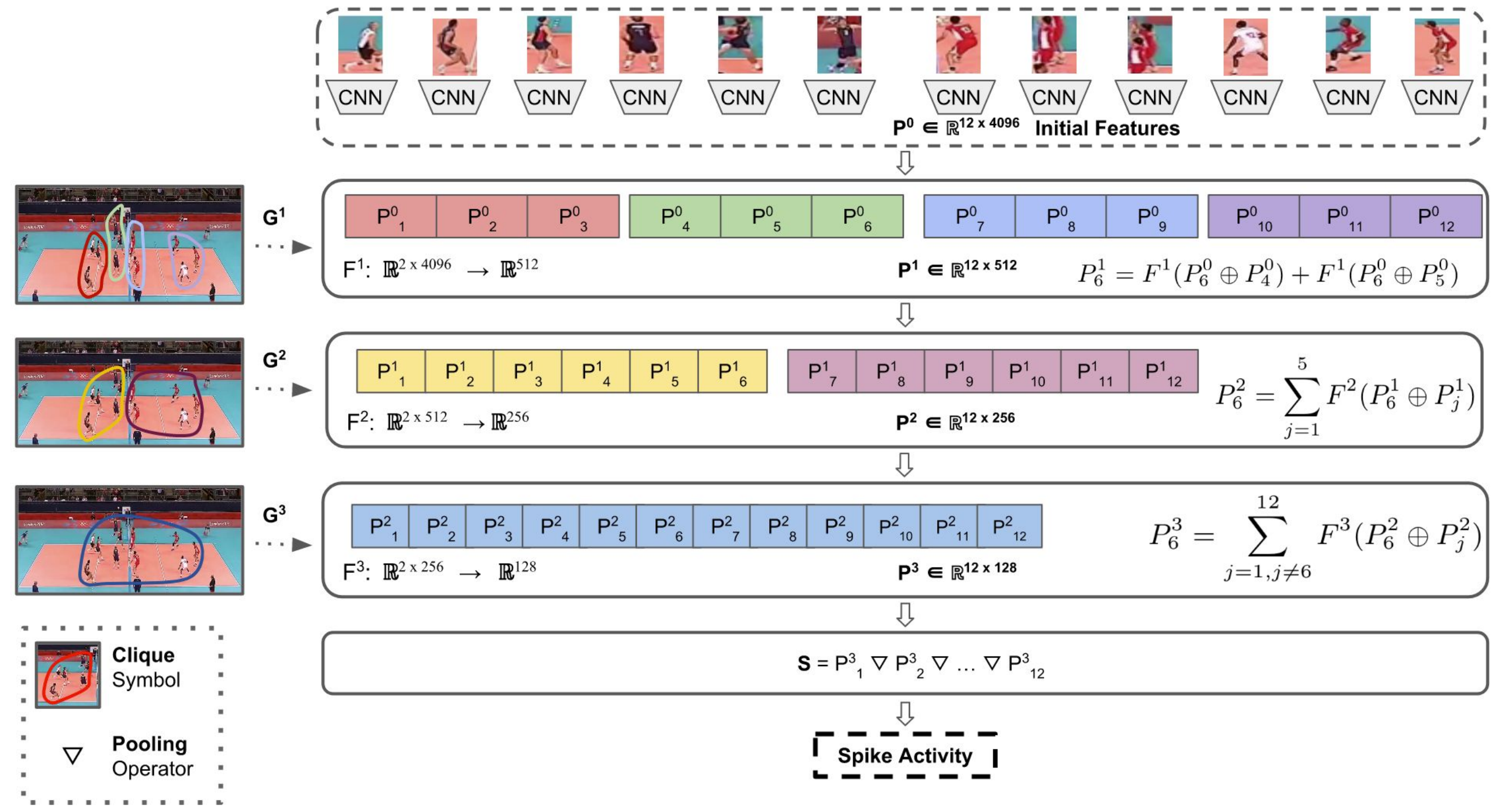
Relational Layer



Top left person has 3 neighbours

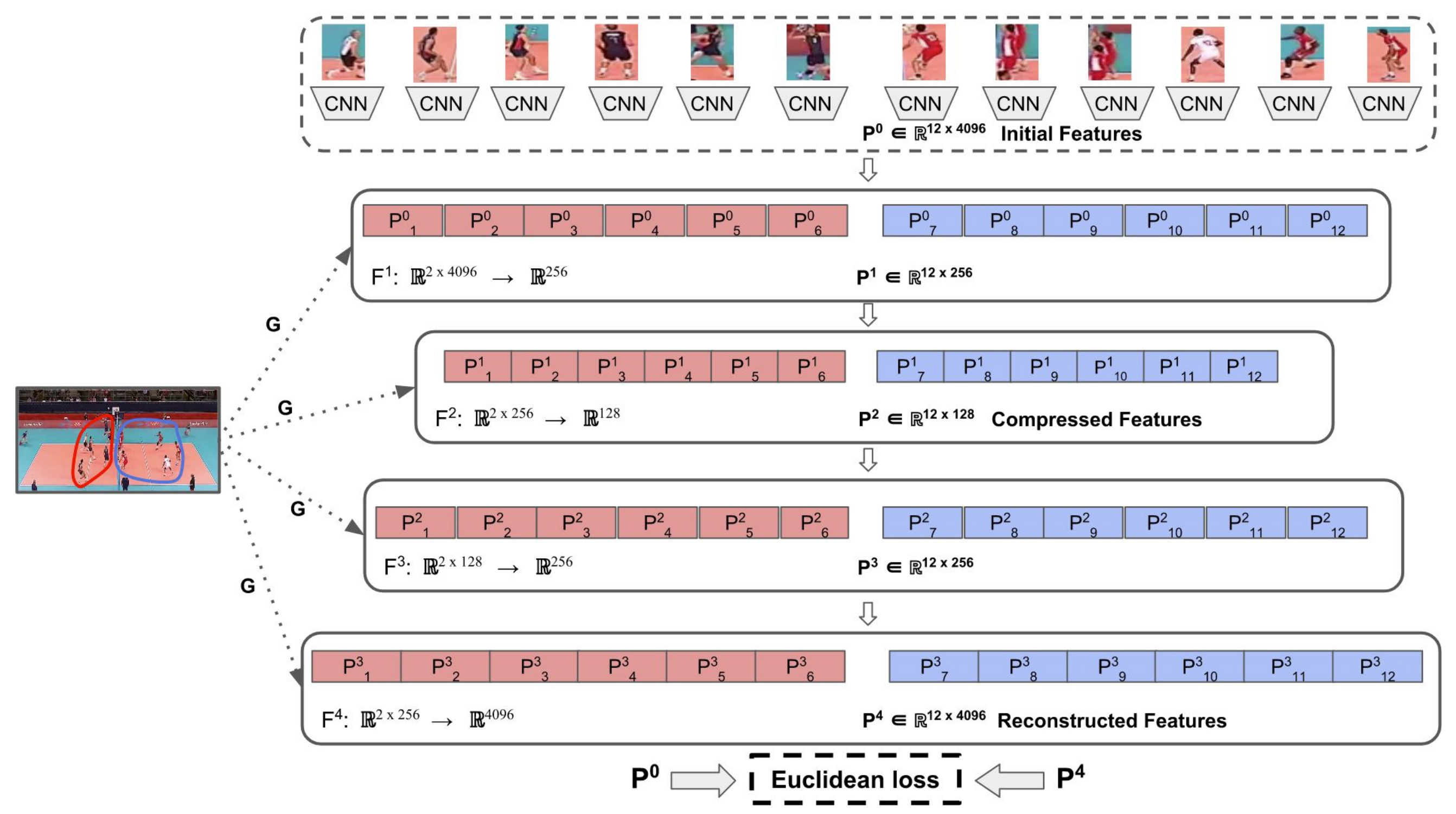
F, an MLP, infers relational edge representation for each neighbour

Hierarchical Relational Network



CNN
$P^0 \in \mathbb{R}^{12 \times 4096}$ Initial Features
G^1
F^1: $\mathbb{R}^{2 \times 4096} \rightarrow \mathbb{R}^{512}$
$P^1 \in \mathbb{R}^{12 \times 512}$
$P^1_6 = F^1(P^0_6 \oplus P^0_4) + F^1(P^0_6 \oplus P^0_5)$
G^2
F^2: $\mathbb{R}^{2 \times 512} \rightarrow \mathbb{R}^{256}$
$P^2 \in \mathbb{R}^{12 \times 256}$
$P^2_6 = \sum_{j=1}^{5} F^2(P^1_6 \oplus P^1_j)$
G^3
F^3: $\mathbb{R}^{2 \times 256} \rightarrow \mathbb{R}^{128}$
$P^3 \in \mathbb{R}^{12 \times 128}$
$P^3_6 = \sum_{j=1, j \neq 6}^{12} F^3(P^2_6 \oplus P^2_j)$
$S = P^3_1 \triangledown P^3_2 \triangledown \dots \triangledown P^3_{12}$
Spike Activity
Clique Symbol
▽ Pooling Operator

Relational AutoEncoder Network



CNN
$P^0 \in \mathbb{R}^{12 \times 4096}$ Initial Features
G
F^1: $\mathbb{R}^{2 \times 4096} \rightarrow \mathbb{R}^{256}$
$P^1 \in \mathbb{R}^{12 \times 256}$
F^2: $\mathbb{R}^{2 \times 256} \rightarrow \mathbb{R}^{128}$
$P^2 \in \mathbb{R}^{12 \times 128}$ Compressed Features
F^3: $\mathbb{R}^{2 \times 128} \rightarrow \mathbb{R}^{256}$
$P^3 \in \mathbb{R}^{12 \times 256}$
F^4: $\mathbb{R}^{2 \times 256} \rightarrow \mathbb{R}^{4096}$
$P^4 \in \mathbb{R}^{12 \times 4096}$ Reconstructed Features
P^0 Euclidean loss P^4

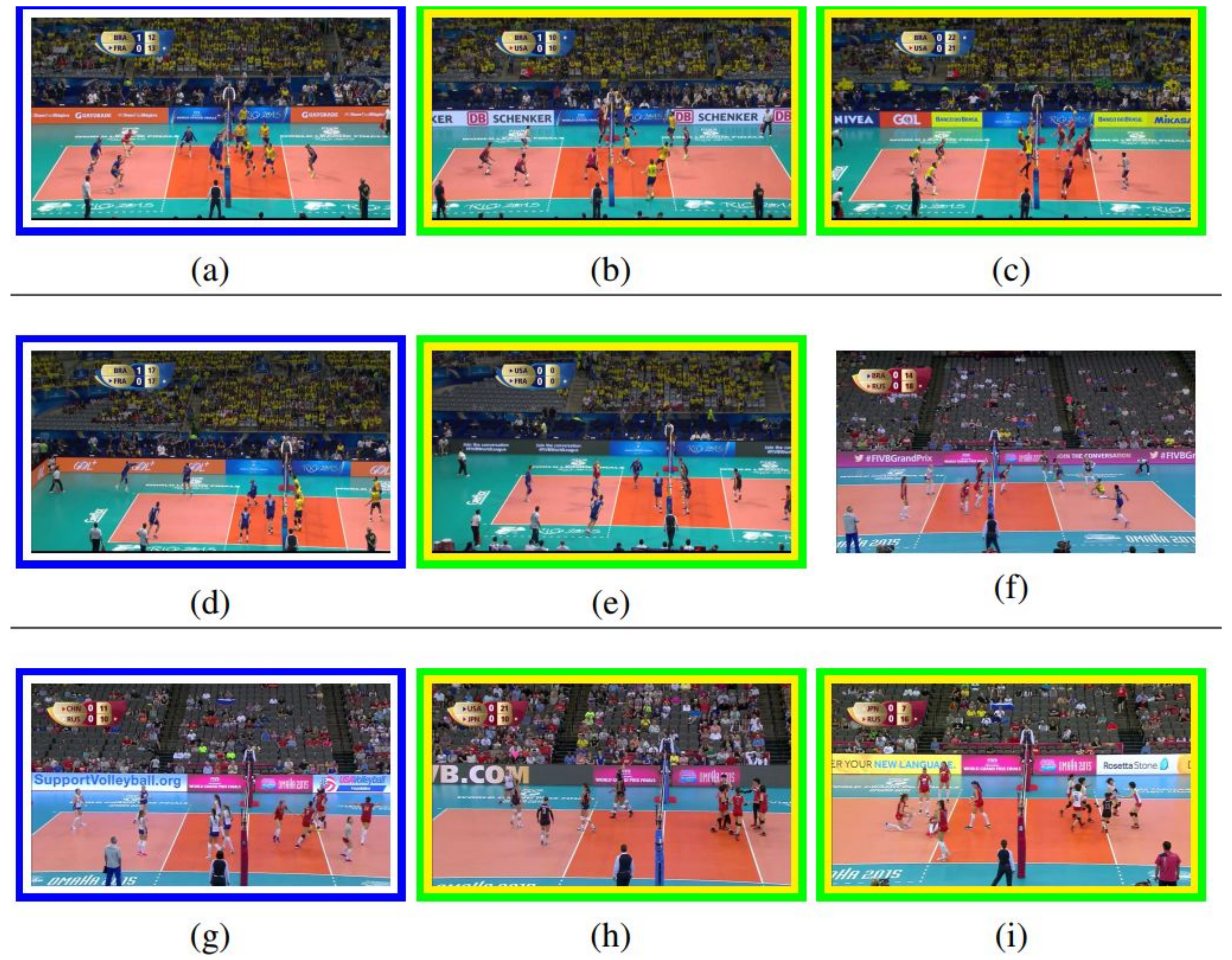
Volleyball Dataset Experiments and Visualizations

Group Activity Recognition Accuracy.
Left table for single step and right table for 10-timesteps

- **2R-21C** = 2 relational layers: First layer has a graph of 2 cliques (one per team). Second layer has a 1 clique graph for all persons in the scene (fully connected graph).

Method	Accuracy
B1-NoRelations	85.1
RCRG-1R-1C	86.5
RCRG-1R-1C-!tuned	75.4
RCRG-2R-11C	86.1
RCRG-2R-21C	87.2
RCRG-3R-421C	86.4
RCRG-2R-11C-conc	**88.3**
RCRG-2R-21C-conc	86.7
RCRG-3R-421C-conc	87.3
Bagautdinov et al. [1]-single	83.8

Method	Accuracy
Bagautdinov et al. [1]	**90.6**
RCRG-2R-11C-conc	89.5
RCRG-2R-21C	89.4
Shu et al. [25]	83.3
Ibrahim et al. [10]	81.9



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Code https://github.com/mostafa-saad/hierarchical-relational-network

Scene retrieval compared to model variants

Method	Hit@1	Hit@2	Hit@3	Hit@4	Hit@5	mAP
RAER-2L-11C	56.8	74.9	84.5	89.8	92.6	**36.8**
RAER-2L-22C	56.9	75.6	84.9	90.0	**93.3**	36.7
RAER-4L-4224C	55.8	76.1	84.0	88.9	92.7	36.6
RAER-4L-2222C	**57.4**	**76.7**	**85.3**	**90.4**	**93.3**	**36.8**

Scene retrieval compared to baselines

Method	Hit@1	Hit@2	Hit@3	Hit@4	Hit@5	mAP
B1-Compact128	49.4	68.7	80.4	87.7	91.4	35.4
B2-VGG19	55.0	73.9	82.7	87.5	91.5	36.4
RAER-4L-2222C	**57.4**	**76.7**	**85.3**	**90.4**	**93.3**	**36.8**

Person retrieval compared to baselines

Method	Hit@1	Hit@2	Hit@3	Hit@4	Hit@5	mAP
B1-Compact128-P	37.7	54.7	64.6	71.7	76.4	22.8
B2-VGG19-P	**47.3**	**63.2**	**72.1**	**77.4**	**81.2**	25.4
RAER-2L-11C-P	45.5	62.2	70.9	76.1	80.1	**25.8**
RAER-4L-2222C-P	42.6	58.3	68.3	73.7	77.8	25.2

Visual Scene retrieval using relational autoencoder

- First blue box is the query image
- Followed by the closest 2 retrievals.
- Green-framed boxes are correct matches