FlowCaps: Optical Flow Estimation with Capsule Networks For Action Recognition

Vinoj Jayasundara
Debaditya Roy
Basura Fernando
Outline

1. Overview
2. Key Contributions
3. FlowCaps: Architecture
4. Key Approaches
5. Experiments and Results
6. Capabilities of FlowCaps
Overview: The need for a Capsule Encoder

Observation: Raw pixels contain sparse motion information, cluttered with non-motion information.

Similar pixel intensities yet differing relative motion. It is convenient for the optical flow estimation if motion information are unentangled and better-coded.
Overview: The need for a Capsule Encoder

Observation: Raw pixels contain sparse motion information, cluttered with non-motion information.

Potential Solution: A capsule encoder, which provides the following:

a) better correspondence matching via finer-grained, concise, motion-specific, and more-interpretable encoding crucial for optical flow estimation
b) better-generalizable optical flow estimation
c) utilize lesser ground truth data
d) significantly reduce the computational complexity

In comparison to the convolutional encoder in FlowNet.
Key Contributions

- Proposing a novel CapsNet based architecture, termed FlowCaps.
- Investigating two contrasting approaches for optical flow estimation and action recognition, namely, frame-wise and segment-wise.
- Achieving a significant (94%) reduction in computational complexity with FlowCaps, in comparison to FlowNet.
- Achieving better optical flow estimation and subsequent action recognition performance for several benchmark datasets.
- Investigating the capabilities of Flow-Caps in terms of out-of-domain generalization and training with only a few samples.
FlowCaps: Architecture
Key Approaches: Improvements to Loss

- Issues with EPE:
  - Only considers the magnitude component in its calculations
  - L2 norm is highly susceptible to outliers with higher values

- We propose:
  \[ L = L_{mag} + \alpha L_{ang} \]

Where \( \alpha \) is an empirically determined constant.
We consider two different approaches based on the number of consecutive frames (k) considered for prediction at a time.

a) Frame-wise (k=2) \( X_{frm} \in \mathbb{R}^{(H \times W \times 2C)} \rightarrow Y_{frm} \in \mathbb{R}^{(H \times W \times 2)} \)

b) Segment-wise (k>2) \( X_{seg} \in \mathbb{R}^{(k \times H \times W \times C)} \rightarrow Y_{seg} \in \mathbb{R}^{(H \times W \times 2)} \)

Intuition behind Segment-wise approach

- The model can benefit from the additional contextual information provided by the extra frames considered.
- In a setting where optical flow estimation and action recognition are performed in tandem, it is natural to consider segments, rather than pairs of frames.
# Results: Optical Flow Estimation

| Model             | EpicFlow [25] | FlowFields [1] | Sintel clean | Sintel final | KITTI15 |
|-------------------|---------------|----------------|--------------|--------------|---------|
| Conventional      |               |                | 2.27         | 3.56         | 9.27    |
|                   | -             | -              | **1.86**     | 3.06         | 8.33    |
| Heavyweight CNN   | FlowNetS [6]  | 38.68          | 4.50         | 5.45         | -       |
|                   | FlowNet2 [17] | 162.49         | 2.02         | 3.54         | 10.08   |
| Lightweight CNN   | LiteFlowNet [16] | 5.37         | 2.48         | 4.04         | 10.39   |
|                   | SPyNet [24]   | 1.20           | 4.12         | 5.57         | -       |
|                   | Ours          | 2.39           | 2.13         | **2.51**     | **7.83** |
## Results: Segment-wise vs Frame-wise

| Model       | KTH-I Frames | Sub UCF-I Frames | UTI-P Frames |
|-------------|--------------|------------------|--------------|
|             | Optical flow estimation performance in EPE | Action classification performance | |
|             | Frame | Seg. | Frame | Seg. | Frame | Seg. | Frame | Seg. | Frame | Seg. | Frame | Seg. |
| FlowNetS    | 1.1934 | 1.1355 | 2.3149 | 2.3079 | 0.4426 | 0.4265 |
| FlowCaps-S  | 1.1033 | **0.9384** | 2.2037 | **2.1930** | 0.3806 | **0.3672** |
| FlowNetS    | 61.30%  | 66.30%  | 85.50%  | 89.70%  | 84.12%  | 83.08%  |
| FlowCaps-S  | **65.00%** | **72.50%** | **91.20%** | **92.30%** | **86.02%** | **85.93%** |
| GT          | 68.90%  |         | 92.60%  |         | 81.37%  |
Results: Optical Flow Estimation and Action Recognition

| Model          | UCF I-Frames | UTI P-Frames | KTH I-Frames | JHMDB  |
|----------------|--------------|--------------|--------------|--------|
|                | test epe     | action       | test epe     | action | test epe     | action       | test epe     | action       |
| GT             | -            | 79.4%        | -            | 81.37% | -            | 68.90%       | -            | 51.49%       |
| FlowNetS       | 1.53         | 55.58%       | 0.44         | 84.12% | 1.19         | 61.30%       | 0.49         | 44.03%       |
| LiteFlowNet    | -            | -            | -            | 83.17% | -            | 59.79%       | -            | 40.30%       |
| SPyNet         | 1.37         | 65.78%       | 0.42         | 87.66% | 0.95         | 64.30%       | 0.44         | 42.54%       |
| Ours           | 1.49         | 64.49%       | 0.39         | 86.02% | 1.10         | 65.00%       | **0.40**     | **48.51%**   |
| Ours - Mod Loss* | 1.41       | -            | 0.35         | -      | 1.04         | -            | 0.26         | -            |
| Ours - Segment | 1.40         | 65.16%       | **0.37**     | 88.34% | **0.93**     | **72.50%**   | 0.71         | 41.90%       |
Optical Flow Estimation: UTI

| Image | Ground Truth | FlowNetS | FlowCaps-S |
|-------|--------------|----------|------------|
| ![Image](image1.png) | ![Ground Truth](ground_truth1.png) | ![FlowNetS](flownets1.png) | ![FlowCaps-S](flowcaps-s1.png) |
| ![Image](image2.png) | ![Ground Truth](ground_truth2.png) | ![FlowNetS](flownets2.png) | ![FlowCaps-S](flowcaps-s2.png) |
| ![Image](image3.png) | ![Ground Truth](ground_truth3.png) | ![FlowNetS](flownets3.png) | ![FlowCaps-S](flowcaps-s3.png) |
Optical Flow Estimation: KTH
Optical Flow Estimation: UCF
FlowCaps: Out-of-Domain Generalization

- We test on all the classes of UCF-101 except for classes with no videos containing more than 5 I-frames, and for the five classes considered for training, which yields 88 out-of-domain action classes.
FlowCaps: Out-of-Domain Generalization
FlowCaps: Training with few samples

- Lower the availability of training data, higher the relative generalization capability of FlowCaps-S.
Thank You!

For more information, please join the Q&A session for the paper ID: 975!

A copy of our paper can be found here: