Learning to Match Aerial Images with Deep Attentive Architectures

Cornell University Department of Computer Science

Georgia Tech College of Computing

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Hani Altwaijry[1,2] Eduard Trulls[3] James Hays[4] Pascal Fua[3] Serge Belongie[1,2]

[1] Department of Computer Science, Cornell University [2] Cornell Tech [3] Computer Vision Laboratory, École Polytechnique Fédérale de Lausanne (EPFL) [4] School of Interactive Computing, College of Computing, Georgia Institute of Technology

Motivation

- Matching **ultra wide-baseline aerial images** goes beyond the reach of traditional tools such as SIFT+RANSAC.
- We approach it with **deep networks** in a classification framework, and obtain **state of the art results.**
- However: can we put **geometry** back into the mix?

Contributions

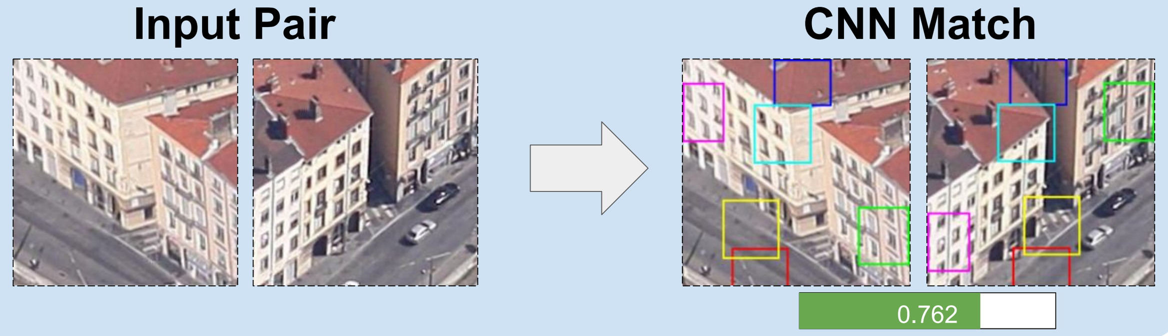
1. We demonstrate that **deep learning offers a solution** for ultra-wide baseline matching.
2. We propose a model that relies on spatial transformers to produce patch matching proposals. We show that **incorporating geometry increases performance.**
3. We conduct a **human study** as a baseline.

Aerial datasets

- "GMaps": Ultra-wide 49k pairs from Google Maps, 3 cities.

- "Lausanne": Wide-baseline 10k pairs from SfM:

Sample Result



Human Performance

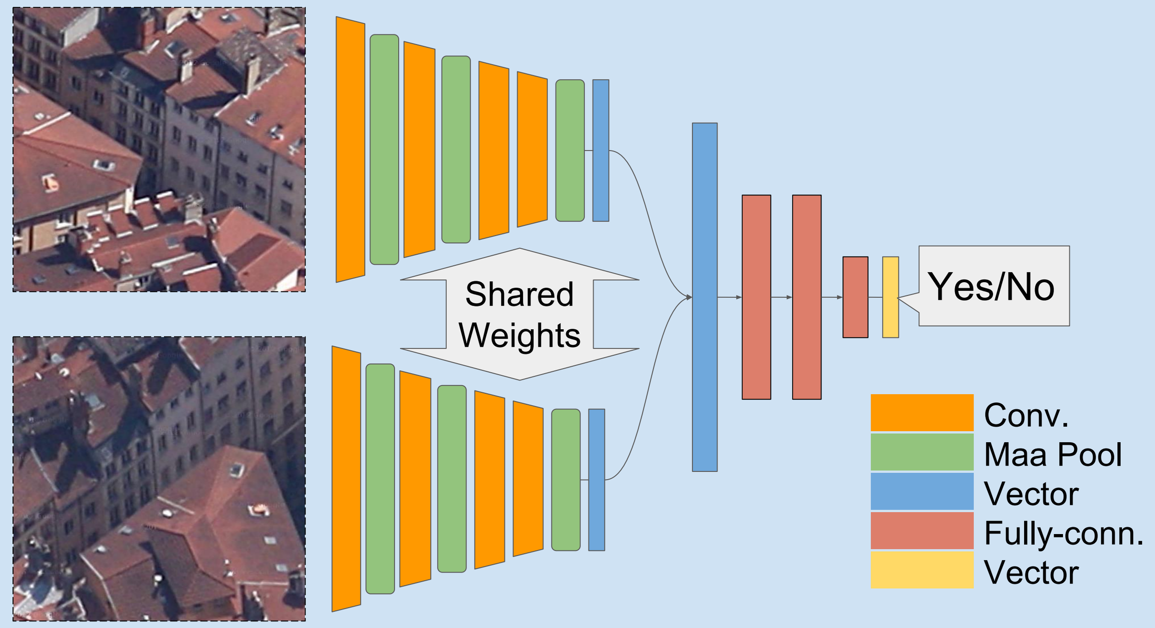
- 1k pairs from the "GMaps" set. Task: Yes/No matching.
- Each pair was shown to 5 participants.
- Results: 93.3% accuracy, 98% precision.

False-Positive

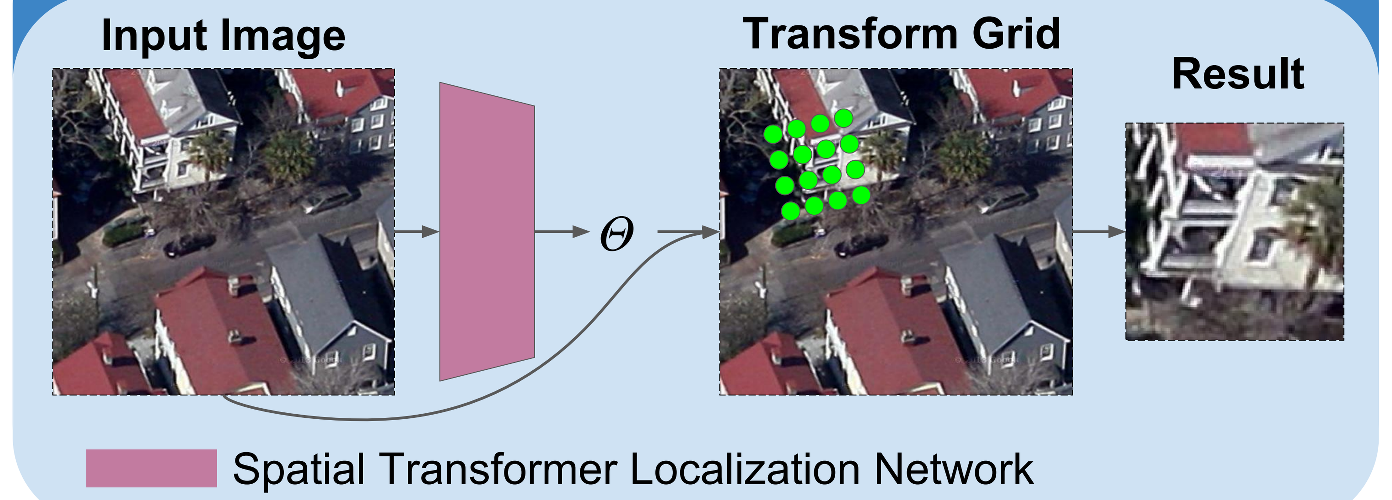
False-Negative

"Hybrid" Model

- Siamese network, with a fine-tuned AlexNet and a matching classifier.
  - ✓ Allows both images to be considered jointly.
  - ✓ Good classification results.
  - ✗ Does not explain why the pair matches or not.

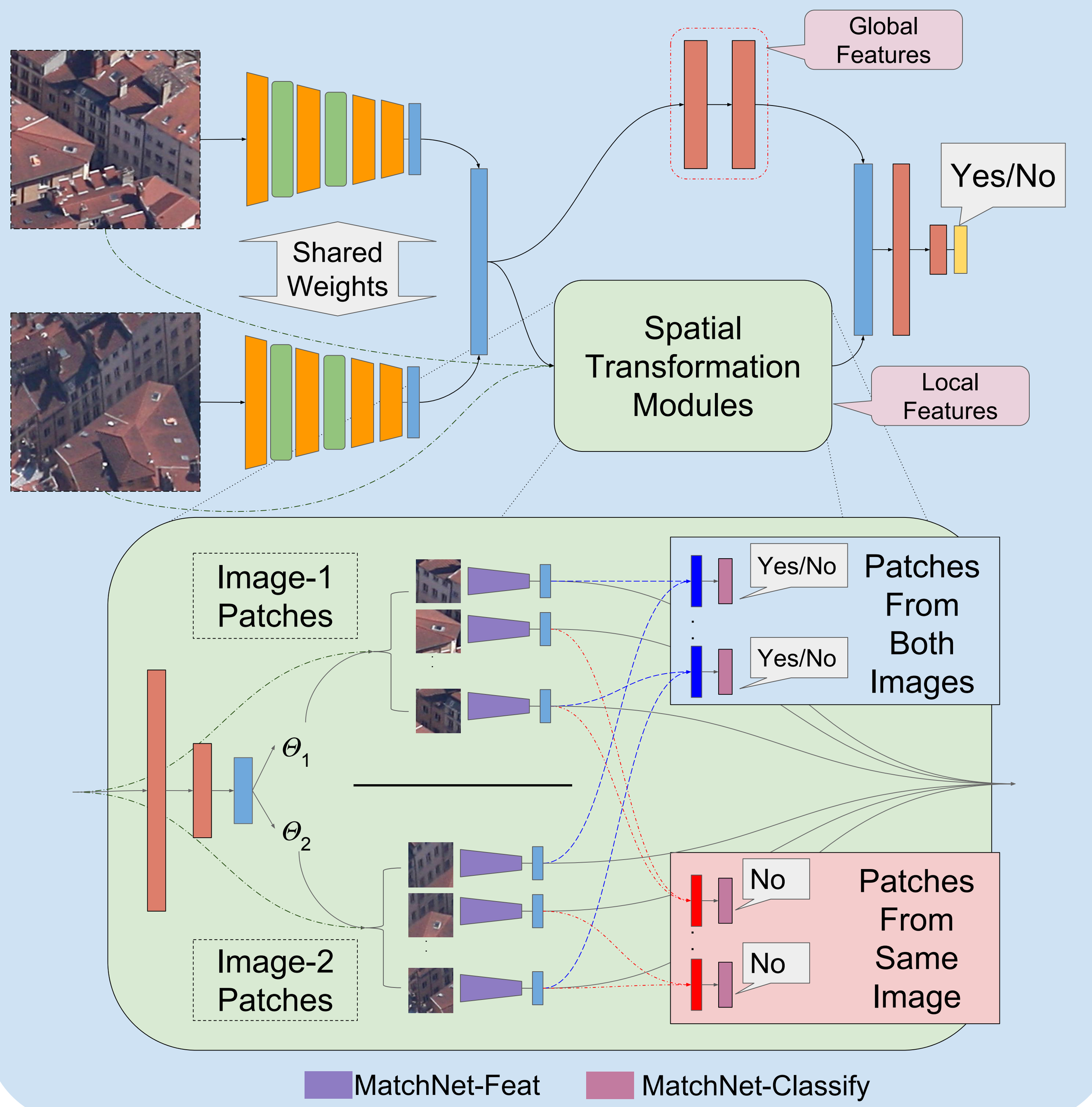


Spatial Transformers



"Hybrid++" model

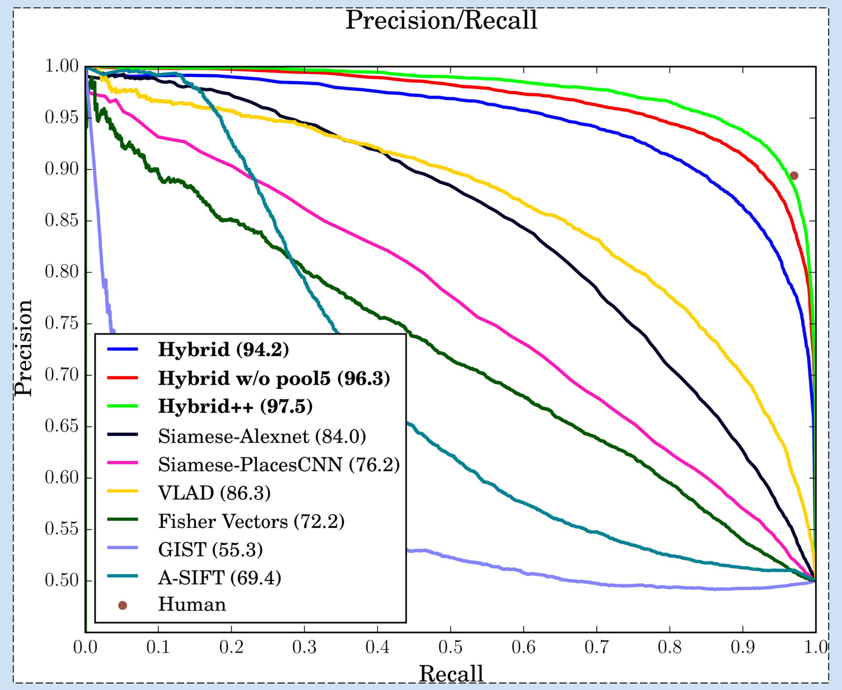
- Compares a pair of images by extracting features globally and locally using spatial transformer modules.
  - ✓ Attempts to explain why the images match.
  - ✓ Specifically models local features.
  - ✓ Local features are extracted given both input images
  - ✓ Jointly trains both global and local features.



Evaluation

Experiments on the "GMaps" Dataset

- 10K Test Pairs, 1:1 positive/negative ratio.
- Viewpoint variations in dataset are small.



Method	Acc.	Acc. pos	Acc. neg	AP
Human*	.933	.894	.972	—
A-SIFT	.613	.353	.874	.694
GIST	.549	.242	.821	.553
Fisher Vectors	.659	.605	.713	.722
VLAD	.786	.769	.803	.863
Siamese PlacesCNN	.690	.626	.754	.762
Siamese AlexNet	.754	.697	.811	.840
Hybrid CNN	.881	.901	.861	.942
Hybrid w/o pool5	.909	.928	.891	.963
Hybrid++	.926	.927	.925	.975

- Sample Matching Results

What is the Spatial Transformers learning? Experiments on "Lausanne"

- Arbitrary viewpoints, with smaller baselines.

Method	Acc.	Acc. pos	Acc. neg	AP
A-SIFT	.947	.896	.998	.968
GIST	.856	.798	.914	.937
Fisher Vectors	.769	.723	.816	.867
VLAD	.898	.867	.930	.965
Siamese PlacesCNN	.690	.626	.754	.958
Siamese AlexNet	.754	.697	.811	.968
Hybrid CNN	.959	.960	.957	.992
Hybrid++	.959	.962	.956	.992

Takeaways

1. Joint-training required individual pre-training of network parts.
2. The spatial transformer is capable of learning varying viewpoint changes per the data.
3. Matches here are not geometric "correspondences", however, we are one step closer.