INRIA and Xerox @ FGComp’13

Philippe-Henri Gosselin (INRIA, ENSEA)
Naila Murray (Xerox)
Hervé Jégou (INRIA)
Florent Perronnin (Xerox)
presented by Giorgos Tolias (INRIA)

FGComp workshop @ ICCV’13
Dec 7, 2013
The system in a nutshell

Based on Fisher vector descriptors:

- extract low-level descriptors (e.g. SIFT) from small patches
- measure “deviation” with respect to a probabilistic model, e.g. a GMM

→ aggregates first and second order statistics

\[
G_{\mu,i}^X = \frac{1}{T \sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left( \frac{x_t - \mu_i}{\sigma_i} \right)
\]

\[
G_{\sigma,i}^X = \frac{1}{T \sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) \left[ \frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right]
\]
The system in a nutshell

Based on Fisher vector descriptors:
- extract low-level descriptors (e.g. SIFT) from small patches
- measure “deviation” with respect to a probabilistic model, e.g. a GMM

→ aggregates first and second order statistics

\[
G_{\mu,i}^X = \frac{1}{T \sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left( \frac{x_t - \mu_i}{\sigma_i} \right)
\]

\[
G_{\sigma,i}^X = \frac{1}{T \sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) \left[ \frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right]
\]

+ linear classifiers
The system in a nutshell

Based on Fisher vector descriptors:
- extract low-level descriptors (e.g. SIFT) from small patches
- measure “deviation” with respect to a probabilistic model, e.g. a GMM

→ aggregates first and second order statistics

\[ G_{\mu,i}^X = \frac{1}{T \sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left( \frac{x_t - \mu_i}{\sigma_i} \right) \]

\[ G_{\sigma,i}^X = \frac{1}{T \sqrt{2 w_i}} \sum_{t=1}^{T} \gamma_t(i) \left[ \frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right] \]

+ linear classifiers

So, what is new?
Our “gold standard” on PASCAL VOC

Our “gold standard” on PASCAL VOC


- images resized to fixed-size
Our “gold standard” on PASCAL VOC


- images resized to fixed-size
- dense descriptors (keep all patches)

- images resized to fixed-size
- dense descriptors (keep all patches)
- SIFT (+ color)
Our “gold standard” on PASCAL VOC


- images resized to fixed-size
- dense descriptors (keep all patches)
- SIFT (+ color)
- small number of Gaussians: 256 by default
Our “gold standard” on PASCAL VOC


- images resized to fixed-size
- dense descriptors (keep all patches)
- SIFT ( + color)
- small number of Gaussians: 256 by default
- spatial pyramids to model scene layout
Our “gold standard” on PASCAL VOC


- images resized to fixed-size
- dense descriptors (keep all patches)
- SIFT (+ color)
- small number of Gaussians: 256 by default
- spatial pyramids to model scene layout
- signed square-root normalization
Our “gold standard” on PASCAL VOC


- images resized to fixed-size
- dense descriptors (keep all patches)
- SIFT (+ color)
- small number of Gaussians: \( k=256 \) by default
- spatial pyramids to model scene layout
- signed square-root normalization

→ focus presentation on differences
Two subsystems, both based on Fisher vector

- Our main submission consists of two subsystems $S_A$ and $S_B$
  - Also evaluated independently (See official results)

- Subsystem $S_A$
  - Almost identical to our standard Fisher classification pipeline
  - Only adjust some important parameters

- Subsystem $S_B$ is designed such that
  - It is as complementary as possible with $S_A$
  - Optimized towards particular domains (cars, aircraft, shoes)
    → Include a few specific processing steps
    → Per-domain Cross-validation of certain parameters (usually fixed)
Good practice #1

Large vocabularies are crucial for fine-grained classification.

Motivation: fine-grain classification seen as in between

- Image classification $\rightarrow k=256$
- Particular object recognition $\rightarrow k=10,000$ to 1 million (with BoW)

![Graph showing mean accuracy vs vocabulary size for different categories with BoW and SIFT-Filter.](image-url)
Good practice #2

Level of details is domain-dependent

→ multi-scale insufficient
→ select size to which an image is resized

<table>
<thead>
<tr>
<th>domain</th>
<th>100k</th>
<th>300k</th>
</tr>
</thead>
<tbody>
<tr>
<td>aircrafts</td>
<td>0.635</td>
<td>0.668</td>
</tr>
<tr>
<td>birds</td>
<td>0.266</td>
<td>0.293</td>
</tr>
<tr>
<td>cars</td>
<td><strong>0.603</strong></td>
<td>0.565</td>
</tr>
<tr>
<td>shoes</td>
<td>0.839</td>
<td><strong>0.862</strong></td>
</tr>
</tbody>
</table>

Remark: fix 100kilo-pixels for dogs for computational reasons
Good practice # 3

Encoding patch location in local descriptors is an alternative to spatial pyramids

Large vocabularies = high dimensional vectors
- Critical issue with spatial pyramid

Subsystem $S_A$
- we use $1\times 1 + 3\times 1$ (instead of $1\times 1 + 3\times 1 + 2\times 2$)
  → give more room, comparatively, to increase $k$

Subsystem $S_B$
- Encode the patch location jointly with the descriptor [Koniuz et al., 13]
  - Does not increase dimensionality (due to PCA to 80 components)
  → allow us to use much larger values of $k$ (up to $k=4,096$)
Good practice # 4

Cross-validating the power normalization is crucial for good performance especially for color descriptor.

\[ S_A, \text{Track 2, domain: birds} \]

- SIFT
- X-color

\[ \alpha \]

\[ \text{mean accuracy} \]
Good practice # 5

Adopt methods that are useful for a few domain only (and let the fusion step decides per domain)

- For instance: filter low-energy patches
  → Interesting for complexity
Filter dense patches
τ=0
\( \tau = 100 \)
τ = 300
τ=400
\[ \tau = 500 \]
τ = 600
\( \tau = 700 \)
τ = 900
τ=1000
Good practice # 5

Adopt methods that are useful for a few domain only (and let the fusion step decides per domain)

- For instance: filter low-energy patches

![Graph showing mean accuracy relative improvement with filtering threshold τ for different categories. Track 1, k=64.]

- Interesting for aircrafts and cars

---

Note: The graph shows the mean accuracy relative improvement with varying filtering thresholds for different categories: aircrafts, birds, cars, dogs, and shoes. The results are labeled as Track 1, k=64.
Good practice # 6

When faced with alternatives design choices, do not choose: combine

$S_A$ and $S_B$ make distinct choices in most steps

- Resolution, sampling, PCA size, descriptor post-processing
- Spatial information
- Classifier
- etc

<table>
<thead>
<tr>
<th>Subsystem</th>
<th>$S_A$</th>
<th>$S_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image (re-)sizing</td>
<td>100k pixels</td>
<td>100k–300k pixels</td>
</tr>
<tr>
<td>Dense sampling</td>
<td>every 4 pixels (x &amp; y)</td>
<td>every 3 pixels</td>
</tr>
<tr>
<td>Input descriptor</td>
<td>SIFT+X-color</td>
<td>SIFT</td>
</tr>
<tr>
<td>Descriptor post-processing</td>
<td>PCA reduction to 64 components</td>
<td>PCA to 80 components</td>
</tr>
<tr>
<td></td>
<td>$x_i := \log(1 + x_i)$</td>
<td>RootSIFT</td>
</tr>
<tr>
<td></td>
<td>X-color: no post-processing</td>
<td>filter low-energy patches ($\tau = 0 \ldots 700$)</td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>1024 − 4096</td>
</tr>
<tr>
<td></td>
<td>spatial pyramid: 1 + 3 × 1</td>
<td>spatial coordinate coding</td>
</tr>
<tr>
<td></td>
<td>Stochastic Gradient Descent</td>
<td>LASVM, $C = 100$</td>
</tr>
</tbody>
</table>
Late fusion

- $S_A$: cross-validation of weights given to SIFT and X-color
  - Remark: already done in previous systems by Xerox
  - Best choices heavily depend on the domain:

![Bar chart showing SIFT vs X-color for different categories: Shoes, Dogs, Cars, Birds, Aircrafts.](chart.png)
Late fusion

- Cross-validation of respective importance of $S_A$ and $S_B$

- $S_B$ receives more importance for domains where filtering and large vocabularies are effective
## Analysis of official FGCOMP’s’s results

### Track 1

<table>
<thead>
<tr>
<th>Team</th>
<th>Aircraft</th>
<th>Birds</th>
<th>Cars</th>
<th>Dogs</th>
<th>Shoes</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inria-Xerox</td>
<td>81.4581</td>
<td>71.6931</td>
<td>87.7876</td>
<td>52.9</td>
<td>91.517</td>
<td>77.0712</td>
</tr>
<tr>
<td>CafeNet*</td>
<td>78.8479</td>
<td>73.0085</td>
<td>79.5797</td>
<td>57.5333</td>
<td>90.1198</td>
<td>75.8178</td>
</tr>
<tr>
<td>Inria-Xerox</td>
<td>75.8776</td>
<td>66.285</td>
<td>84.7034</td>
<td>50.4167</td>
<td>88.6228</td>
<td>73.1811</td>
</tr>
<tr>
<td>VisionMetric*</td>
<td>75.4875</td>
<td>63.8977</td>
<td>74.3316</td>
<td>55.8667</td>
<td>89.022</td>
<td>71.7211</td>
</tr>
<tr>
<td>Symbiotic</td>
<td>75.8476</td>
<td>69.0621</td>
<td>81.0347</td>
<td>44.8917</td>
<td>87.3253</td>
<td>71.6323</td>
</tr>
<tr>
<td>Inria-Xerox</td>
<td>80.5881</td>
<td>58.5384</td>
<td>84.6661</td>
<td>35.6167</td>
<td>90.9182</td>
<td>70.0655</td>
</tr>
<tr>
<td>CognitiveVision*</td>
<td>67.4167</td>
<td>72.7893</td>
<td>64.395</td>
<td>60.5583</td>
<td>84.8303</td>
<td>69.9979</td>
</tr>
<tr>
<td>DPD_Berkeley*</td>
<td>68.4668</td>
<td>69.5737</td>
<td>67.4046</td>
<td>50.8417</td>
<td>89.521</td>
<td>69.1615</td>
</tr>
<tr>
<td>VisionMetric</td>
<td>73.9274</td>
<td>51.352</td>
<td>69.3073</td>
<td>38.6333</td>
<td>87.3253</td>
<td>64.1091</td>
</tr>
<tr>
<td>CognitiveVision</td>
<td>58.8059</td>
<td>51.6931</td>
<td>52.3691</td>
<td>47.3667</td>
<td>78.1437</td>
<td>57.6757</td>
</tr>
<tr>
<td>MPG</td>
<td>9.45095</td>
<td>54.5676</td>
<td>69.27</td>
<td>42.9167</td>
<td>88.4232</td>
<td>52.9257</td>
</tr>
<tr>
<td>MPG</td>
<td>9.45095</td>
<td>56.4677</td>
<td>63.7732</td>
<td>0.975</td>
<td>88.4232</td>
<td>43.818</td>
</tr>
<tr>
<td>Infor_FG*</td>
<td>30.393</td>
<td>9.06212</td>
<td>4.45218</td>
<td>0.816667</td>
<td>35.2295</td>
<td>15.9907</td>
</tr>
<tr>
<td>InterfAlce</td>
<td>5.79058</td>
<td>2.55786</td>
<td>1.11926</td>
<td>6.95833</td>
<td>5.98802</td>
<td>4.48281</td>
</tr>
</tbody>
</table>

* Indicates using features learned on outside data (e.g. ILSVRC2012)
### Analysis of official FGCOMP’s results

#### Track 1

<table>
<thead>
<tr>
<th>Team</th>
<th>Aircraft</th>
<th>Birds</th>
<th>Cars</th>
<th>Dogs</th>
<th>Shoes</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inria-Xerox</td>
<td>81.4581</td>
<td>71.6931</td>
<td>87.7876</td>
<td>52.9</td>
<td>91.517</td>
<td>77.0712</td>
</tr>
<tr>
<td>CafeNet*</td>
<td>78.8479</td>
<td>73.0085</td>
<td>79.5797</td>
<td>57.5333</td>
<td>90.1198</td>
<td>75.8178</td>
</tr>
<tr>
<td>Inria-Xerox</td>
<td>75.8776</td>
<td>66.285</td>
<td>84.7034</td>
<td>50.4167</td>
<td>88.6228</td>
<td>73.1811</td>
</tr>
<tr>
<td>VisionMetric*</td>
<td>75.4875</td>
<td>63.8977</td>
<td>74.3316</td>
<td>55.8667</td>
<td>89.022</td>
<td>71.7211</td>
</tr>
<tr>
<td>Symbiotic</td>
<td>75.8476</td>
<td>69.0621</td>
<td>81.0347</td>
<td>44.8917</td>
<td>87.3253</td>
<td>71.6323</td>
</tr>
<tr>
<td>Inria-Xerox</td>
<td>80.5881</td>
<td>58.5384</td>
<td>84.6661</td>
<td>35.6167</td>
<td>90.9182</td>
<td>70.0655</td>
</tr>
<tr>
<td>CognitiveVision*</td>
<td>67.4167</td>
<td>72.7893</td>
<td>64.395</td>
<td>60.5583</td>
<td>84.8303</td>
<td>69.9979</td>
</tr>
<tr>
<td>DPD_Berkeley*</td>
<td>68.4668</td>
<td>69.5737</td>
<td>67.4046</td>
<td>50.8417</td>
<td>89.521</td>
<td>69.1615</td>
</tr>
<tr>
<td>VisionMetric</td>
<td>73.9274</td>
<td>51.352</td>
<td>69.3073</td>
<td>38.6333</td>
<td>87.3253</td>
<td>64.1091</td>
</tr>
<tr>
<td>CognitiveVision</td>
<td>58.8059</td>
<td>51.6931</td>
<td>52.3691</td>
<td>47.3667</td>
<td>78.1437</td>
<td>57.6757</td>
</tr>
<tr>
<td>MPG</td>
<td>9.45095</td>
<td>54.5676</td>
<td>69.27</td>
<td>42.9167</td>
<td>88.4232</td>
<td>52.9257</td>
</tr>
<tr>
<td>MPG</td>
<td>9.45095</td>
<td>56.4677</td>
<td>63.7732</td>
<td>0.975</td>
<td>88.4232</td>
<td>43.818</td>
</tr>
<tr>
<td>Infor_FG*</td>
<td>30.393</td>
<td>9.06212</td>
<td>4.45218</td>
<td>0.816667</td>
<td>35.2295</td>
<td>15.9907</td>
</tr>
<tr>
<td>InterfAlce</td>
<td>5.79058</td>
<td>2.55786</td>
<td>1.11926</td>
<td>6.95833</td>
<td>5.98802</td>
<td>4.48281</td>
</tr>
</tbody>
</table>

* Indicates using features learned on outside data (e.g. ILSVRC2012)
Analysis of official FGCOMP's results

**Track 1**

<table>
<thead>
<tr>
<th>Team</th>
<th>Aircraft</th>
<th>Birds</th>
<th>Cars</th>
<th>Dogs</th>
<th>Shoes</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inria-Xerox</td>
<td>81.4581</td>
<td>71.6931</td>
<td>87.7876</td>
<td>52.9</td>
<td>91.517</td>
<td>77.0712</td>
</tr>
<tr>
<td>CafeNet*</td>
<td>78.8479</td>
<td>73.0085</td>
<td>79.5797</td>
<td>57.5333</td>
<td>90.1198</td>
<td>75.8178</td>
</tr>
<tr>
<td>Inria-Xerox</td>
<td>75.8776</td>
<td>66.285</td>
<td>84.7034</td>
<td>50.4167</td>
<td>88.6228</td>
<td>73.1811</td>
</tr>
<tr>
<td>VisionMetric*</td>
<td>75.4875</td>
<td>63.8977</td>
<td>74.3316</td>
<td>55.8667</td>
<td>89.022</td>
<td>71.7211</td>
</tr>
<tr>
<td>Symbiotic</td>
<td>75.8476</td>
<td>69.0621</td>
<td>81.0347</td>
<td>44.8917</td>
<td>87.3253</td>
<td>71.6323</td>
</tr>
<tr>
<td>Inria-Xerox</td>
<td>80.5881</td>
<td>58.5384</td>
<td>84.6661</td>
<td>35.6167</td>
<td>90.9182</td>
<td>70.0655</td>
</tr>
<tr>
<td>CognitiveVision*</td>
<td>67.4167</td>
<td>72.7893</td>
<td>64.395</td>
<td>60.5583</td>
<td>84.8303</td>
<td>69.9979</td>
</tr>
<tr>
<td>DPD_Berkeley*</td>
<td>68.4668</td>
<td>69.5737</td>
<td>67.4046</td>
<td>50.8417</td>
<td>89.521</td>
<td>69.1615</td>
</tr>
<tr>
<td>VisionMetric</td>
<td>73.9274</td>
<td>51.352</td>
<td>69.3073</td>
<td>38.6333</td>
<td>87.3253</td>
<td>64.1091</td>
</tr>
<tr>
<td>CognitiveVision</td>
<td>58.8059</td>
<td>51.6931</td>
<td>52.3691</td>
<td>47.3667</td>
<td>78.1437</td>
<td>57.6757</td>
</tr>
<tr>
<td>MPG</td>
<td>9.45095</td>
<td>54.5676</td>
<td>69.27</td>
<td>42.9167</td>
<td>88.4232</td>
<td>52.9257</td>
</tr>
<tr>
<td>MPG</td>
<td>9.45095</td>
<td>56.4677</td>
<td>63.7732</td>
<td>0.975</td>
<td>88.4232</td>
<td>43.818</td>
</tr>
<tr>
<td>Infor_FG*</td>
<td>30.393</td>
<td>9.06212</td>
<td>4.45218</td>
<td>0.816667</td>
<td>35.2295</td>
<td>15.9907</td>
</tr>
<tr>
<td>InterfAlce</td>
<td>5.79058</td>
<td>2.55786</td>
<td>1.11926</td>
<td>6.95833</td>
<td>5.98802</td>
<td>4.48281</td>
</tr>
</tbody>
</table>

* Indicates using features learned on outside data (e.g. ILSVRC2012)
Analysis of official FGCOMP’s results

Track 2

<table>
<thead>
<tr>
<th>Team</th>
<th>Aircraft</th>
<th>Birds</th>
<th>Cars</th>
<th>Dogs</th>
<th>Shoes</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inria-Xerox</td>
<td><strong>80.7381</strong></td>
<td><strong>49.8173</strong></td>
<td><strong>82.7136</strong></td>
<td><strong>45.7083</strong></td>
<td><strong>88.1238</strong></td>
<td><strong>69.4202</strong></td>
</tr>
<tr>
<td>Symbiotic</td>
<td>72.4872</td>
<td>46.0171</td>
<td>77.9878</td>
<td>37.1417</td>
<td><strong>89.1218</strong></td>
<td>64.5511</td>
</tr>
<tr>
<td>Inria-Xerox</td>
<td><strong>66.3966</strong></td>
<td>44.5067</td>
<td><strong>76.3462</strong></td>
<td><strong>43.9583</strong></td>
<td><strong>86.3273</strong></td>
<td><strong>63.507</strong></td>
</tr>
<tr>
<td>Inria-Xerox</td>
<td><strong>80.7381</strong></td>
<td><strong>34.4458</strong></td>
<td><strong>76.8934</strong></td>
<td><strong>24.4</strong></td>
<td><strong>87.3253</strong></td>
<td><strong>60.7605</strong></td>
</tr>
<tr>
<td>DPD_Berkeley*</td>
<td>45.5146</td>
<td>42.704</td>
<td>43.3777</td>
<td>41.9083</td>
<td>59.98</td>
<td>46.6969</td>
</tr>
<tr>
<td>Infor_FG*</td>
<td>9.66097</td>
<td>5.74909</td>
<td>3.70601</td>
<td>32.7083</td>
<td>4.69062</td>
<td>11.303</td>
</tr>
<tr>
<td>InterfAlce</td>
<td>5.43054</td>
<td>2.58222</td>
<td>1.16901</td>
<td>6.94167</td>
<td>5.28942</td>
<td>4.28257</td>
</tr>
</tbody>
</table>

* Indicates using features learned on outside data (e.g. ILSVRC2012)
Conclusion

• Our participation: guided by our view of FGC
  ► Task as in between image classification and particular object recognition

• Our system: best performer while derived from established techniques
  ► A typical Fisher-based classification system
  ► Another Fisher-based system, but using different design choices in most steps

• Fisher vector is a good contender for the state of the art in FGC
  ► With large vocabularies
  ► With cross-validation, per domain, of the key parameters

• Good strategies depend on the domain
  ► Color is important for birds and dogs
  ► Filtering strategy: performance boost for some domains
    → cross-validate late-fusion to make these choice automatically
    → requires complementary systems
Thanks

This work was partially funded by the French ANR project FIRE-ID
http://fire-id.gforge.inria.fr/

Technical report with more details: available soon!