ECE 5582 Computer Vision
Lec 11: Object Re-Identification

Zhu Li
Dept of CSEE, UMKC
Office: FH560E, Email: lizhu@umkc.edu, Ph: x 2346.
http://l.web.umkc.edu/lizhu
Outline

- ReCap of Lecture 10
  - Logistic Regression
  - SVM

- Visual Object Re-Identification
  - MPEG CDVS Research
  - Google Landmark Challenge

- Summary
Logistic Regression

- Logistic Function:
  - Mapping linear function to [0 1].
  - Give a prob of observed $X$ is has label 0 or 1 via logistic mapping

\[ z = \sum_{i} w_i x_i + w_0 = WX \]

\[ g(z) = \frac{1}{1 - e^z} \]
With Log loss function the problem is convex, and a gradient search solution can reach optimal.

For each weights $w_j$, we have,

$$\frac{\partial}{\partial w_j} L(w) = -(y \frac{1}{g(w^T x)} - (1 - y) \frac{1}{1 - g(w^T x)}) \frac{\partial}{\partial w_j} g(w^T x)$$

$$= -(y \frac{1}{g(w^T x)} - (1 - y) \frac{1}{1 - g(w^T x)}) g(w^T x) (1 - g(w^T x)) \frac{\partial}{\partial w_j} w^T x$$

$$= -(y(1 - g(w^T x)) - (1 - y) g(w^T x)) x_j$$

$$= (h(x) - y) x_j$$

For a batch of $m$ observed $\{x^{(i)}, y^{(i)}\}$, we can do batch descent, or stochastic gradient descent (SGD)

$$\frac{\partial}{\partial w_j} L(w) = \frac{1}{m} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)}) x_j$$

matlab: \texttt{mnrfit()}
ReCap of Lec 10

- Support Vector Machine:
SVM Summary

What is a good classifier?
- Not only good precision-recall performance at training (empirical risk function), but also need to consider the model complexity
- Structural Risk: penalizing by \( VC \) dimension

VC dimension
- Is a good measure of model complexity
- How many data points a certain classifier can shatter, higher the dimension easier to shatter (why kernel is good)

SVM:
- For linear hyperplane decision function, structural risk minimization is equivalent to \( gap \) maximization
- Lagrangian Relaxation & Primal-Dual decomposition
- \textit{Support Vectors}: mathematically, are data points that has non-zero Lagrangian associated with
- \textit{Kernel Trick}: implicit mapping to higher dimensional richer structure space. Heuristic, may have overfitting risks (eg. RBF)
Lagrangian & Primal-Dual Decomposition

- Primal-Dual Decomposition,
  - Let’s write the Lagrangian
    \[
    L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} \alpha_i y_i (w \cdot x_i + b) + \sum_{i=1}^{n} \alpha_i
    \]
  - The derivatives
    \[
    \frac{\partial L(w, b, \alpha)}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i, \quad \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i
    \]
    \[
    \frac{\partial L(w, b, \alpha)}{\partial b} = - \sum_{i=1}^{n} \alpha_i y_i, \quad \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0
    \]
  - So the dual function
    \[
    q(\alpha) = \inf_{w,b} L(w, b, \alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)
    \]
SVM on HoG in Matlab

SVM is only dealing with a 2-class problem

% hogs svm
[n_img, kd]=size(hogs);
hogs_lbl = zeros(1, n_img); hogs_lbl(1:20) = 1;
svm_hogs = svmtrain(hogs, hogs_lbl);
% recog
rec_lbls = svmclassify(svm_hogs, hogs);

fprintf('
 error rate = %1.4f', sum(abs(hogs_lbl - rec_lbls'))/n_img);

stem(abs(hogs_lbl - rec_lbls'), '.'); hold on; grid on;
axis([1 160 0 5]); str=sprintf('err rate=%1.2f ',
sum(abs(hogs_lbl - rec_lbls'))/n_img); title(str);
SVM on HoG Performance

- 4 image sets from CalTech101: against the rest

- 2: Faces_easy
- 3: Leopards
- 26: cougar_face
- 33: dollar_bill
Outline

- ReCap of Lecture 10
  - SVM
- HW-2
- Visual Object Re-Identification
- Summary
Object Re-Id Outline

- The Problem and MPEG CDVS Standardization Scope
- CDVS Query Extraction and Compression Pipeline
  - Key point Detection and Selection (ALP, CABOX, FS)
  - Global Descriptor: Key points aggregation and compression (SCFV, RVD, AKULA)
  - Local Descriptor: Key points and coordinates compression
- CDVS Query Processing
  - Pair-wise Matching
  - Retrieval
  - Indexing
- CDVA Work
  - Handling video input
  - Image/Video Understanding
- Summary
Mobile Visual Search Problem

- CDVS: Object Identification: bridging the real and cyber world

- Image Understanding/Tagging: associate labels with pixels
CDVS Scope

- MPEG CDVS Standardization Scope
  - Define the visual query bit-stream extracted from images
  - Front-end: image feature capture and compression
  - Server Back-end: image feature indexing and query processing

- Objectives/Challenges:
  - **Real-time**: front end real time performance, e.g., 640x480 @30fps
  - **Compression**: Low bit rate over the air, achieving 20 X compression w.r.t to sending images, or 10X compression of the raw features.
  - **Matching Accuracy**: >95% accuracy in pair-wise matching (verification) and >90% precision in identification
  - **Indexing/Search Efficiency**: real time backend response from large (>100m) visual repository
Object Re-Identification via SIFT

What are the problems?

- Accuracy
- Speed
- Query Compression
- Localization
MPEG-7 CDVS, 8th FP7 Networked Media Concentration meeting, Brussels, December 13, 2011. (thanks, Yuriy!)
CDVS Data Set Performance

- **Annotated Data Set:**
  - Mix of graphics, landmarks, buildings, objects, video clips, and paintings.
  - Approx 32k images

- **Distraction Set:**
  - Approx 1m images from various places

- **Image Query Size**
  - 512 bytes to 16K bytes
  - 4k~8k bytes are the most useful
## The (Long)MPEG CDVS Time Line

<table>
<thead>
<tr>
<th>Meeting / Location/ Date</th>
<th>Action</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>96th meeting, Geneva</strong></td>
<td>Final CfP published Available databases and evaluation software.</td>
<td></td>
</tr>
<tr>
<td>Mar 25, 2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>97th meeting, Torino</strong></td>
<td>Last changes in databases and evaluation software.</td>
<td>Dataset: 30k annotated images + 1m distractor images.</td>
</tr>
<tr>
<td>July 18-23, 2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>98th meeting, Geneva</strong></td>
<td>Evaluation of proposals</td>
<td>11 proposals received, selected TI Lab test model under consideration.</td>
</tr>
<tr>
<td>Nov 26 – Dec 2, 2011</td>
<td></td>
<td>Working Draft 1</td>
</tr>
<tr>
<td><strong>99th meeting, San Jose</strong></td>
<td>WD</td>
<td>Working Draft 2</td>
</tr>
<tr>
<td>Feb 10, 2012</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>100th meeting, Geneva</strong></td>
<td>WD</td>
<td>Working Draft 2</td>
</tr>
<tr>
<td>March 4, 2012</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>…..</strong></td>
<td>WD</td>
<td>Working Draft 2</td>
</tr>
<tr>
<td><strong>106th meeting, Geneva</strong></td>
<td>WD</td>
<td>Working Draft 2</td>
</tr>
<tr>
<td>Oct, 2013</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>108th meeting, Valencia</strong></td>
<td>CD</td>
<td>SIFT patent issue</td>
</tr>
<tr>
<td>Mar, 2014</td>
<td></td>
<td>Committee Draft</td>
</tr>
<tr>
<td><strong>108th meeting, Valencia</strong></td>
<td>DIS</td>
<td>Finally….</td>
</tr>
</tbody>
</table>
Outline

The Problem and CDVS Standardization Scope

CDVS Query Extraction and Compression Pipeline
- Key point Detection and Selection (ALP, CABOX, FS)
- Global Descriptor: Key points aggregation and compression (SCFV, RVD, AKULA)
- Local Descriptor: Key points and coordinates compression

CDVS Query Processing
- Pair-wise Matching
- Retrieval
- Indexing

CDVA Work
- Handling video input
- Image/Video Understanding

Summary
CDVS Query Processing Pipeline

- **KD** - Keypoint Detection
  - ALP, BFLoG, CABOX
- **FS** - Feature Selection
- **GD** - Global Descriptor from key points *aggregation*
  - SCFV, RVD, AKULA
- **LD** – Local Descriptor
  - SIFT compression, spatial coordinates compression
How SIFT Works

- **Detection:** find extrema in spatial-scale space
  - Scale space is represented by LoG filtering output, but approximated in SIFT by DoG

- **Key Points Representation:** 128 dim
Keypoint (SIFT) Detection

Main motivation

- Find invariant and repeatable features to identify objects
  - Many options, SURF, SIFT, BRISK, …etc, but SIFT still best performing
- Circumvent the SIFT patent, which was sold to an unknown third company
  - SIFT patent main claim: DoG filtering to reconstruct scale space
- Extraction Speed: can we be faster than DoG?

Main Proposals:

- Telecom Italia: ALP – Scale Space SIFT Detector (adopted !)
- Samsung Research: CABOX – Integral Image Domain Fast Box Filtering (incomplete results)
- Beijing University/ST Micro/Huawei: Freq domain Gaussian Filtering (deemed not departing far enough from the SIFT patent)
Telecom Italia: Gianluca Francini, Skjalg Lepsoy, Massimo Balestri

key idea, model the scale space response as a polynomial function, and estimate its coefficients by LoG filtering at different scales:

\[ h(x, y, \sigma) = \sigma^2 \cdot \left( \frac{d^2}{dx^2} + \frac{d^2}{dy^2} \right) g(x, y, \sigma) \]

\[ h(m, n, \sigma) \approx \sum_{k=1}^{K} \gamma_k(\sigma) \cdot h(m, n, \sigma_k) \]

\[ \gamma_k(\sigma) \approx a_k \sigma^3 + b_k \sigma^2 + c_k \sigma + d_k \]

LoG kernel at any scale can be expressed as l.c. of 4 fixed scale kernels.
Express the scale response at \((x, y)\), as a 3rd order polynomial

\[
p(x, y, \sigma) = \alpha_3(x, y)\sigma^3 + \alpha_2(x, y)\sigma^2 + \alpha_1(x, y)\sigma + \alpha_0(x, y)
\]

The polynomial coefficients are obtained by filtering, where \(L_k\) is the LoG filtered image at pre-fixed scale \(k\):

<table>
<thead>
<tr>
<th>(k)</th>
<th>(a_k)</th>
<th>(b_k)</th>
<th>(c_k)</th>
<th>(d_k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.2464</td>
<td>2.5021</td>
<td>-8.2007</td>
<td>8.6432</td>
</tr>
<tr>
<td>1</td>
<td>0.4934</td>
<td>-4.5636</td>
<td>12.9824</td>
<td>-10.8424</td>
</tr>
<tr>
<td>2</td>
<td>-0.2717</td>
<td>2.0108</td>
<td>-4.0449</td>
<td>2.1204</td>
</tr>
<tr>
<td>3</td>
<td>0.0140</td>
<td>0.1549</td>
<td>-1.0565</td>
<td>1.3886</td>
</tr>
</tbody>
</table>

\(\alpha_3(x, y) = \sum_{k=0}^{K-1} a_k \cdot L_k(x, y)\)

\(\alpha_2(x, y) = \sum_{k=0}^{K-1} b_k \cdot L_k(x, y)\)

\(\alpha_1(x, y) = \sum_{k=0}^{K-1} c_k \cdot L_k(x, y)\)

\(\alpha_0(x, y) = \sum_{k=0}^{K-1} d_k \cdot L_k(x, y)\)
ALP Scale Space Extrema Detection

ALP filtering scale space response as polynomials

- Scale Space Extrema: \((I \ast h)[430, 122, \sigma] \approx 2.85 \sigma^3 - 39.51 \sigma^2 + 172.65 \sigma - 172.61\)

- No Extrema (saddle point): \((I \ast h)[153, 356, \sigma] \approx 0.12 \sigma^3 - 1.06 \sigma^2 + 3.15 \sigma - 2.5\)
Displacement refinement

- Scale response as a 2\textsuperscript{nd} order polynomial function of displacement \((u, v)\)

\[
(I * h)[x - u, y - v, \sigma] \\
\approx \beta_5(x, y, \sigma)u^2 + \beta_4(x, y, \sigma)v^2 + \beta_3(x, y, \sigma)uv + \beta_2(x, y, \sigma)u + \beta_1(x, y, \sigma)v + \beta_0(x, y, \sigma)
\]
Repeatability vs VL_FEAT SIFT:
Gaussian Filter $g_\sigma$

Box Filters $b_\sigma$

Approximate DoG/LoG by a cascade of box filters that can offer early termination:

- Very fast integral image domain box filtering,
- The box filters are found by solving the following problem, sparse combination of box filters, via LASSO:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \| g - Bh \|_2^2 + \lambda \| h \|_1 \\
\text{subject to} & \quad 1^T Bh = \alpha
\end{align*}
\]
residual = \|g - Bh\|_2

The influence of the used dictionary determines not only the quality of the approximation but also the number of boxes required.
More examples of keypoint detection using box filters.

Overlap: 85%

Overlap: 88%

- Algorithmically the fastest SIFT detector amongst CE1 contributions
- Ref: V. Fragoso, G. Srivastava, A. Nagar, Z. Li, K. Park, and M. Turk, "Cascade of Box (CABOX) Filters for Optimal Scale Space Approximation", *Proc of the 4th IEEE Int'l Workshop on Mobile Vision*, Columbus, USA, 2014
Feature Selection

Why do Feature Selection?

- Average 1000+ SIFTs extracted for VGA sized images, need to reduce the number of actual SIFTs sent
- Not all SIFTs are created equal in repeatability in image match,
  - model the repeatability as a prob function [Lepsøy, S., Francini, G., Cordara, G., & Gusmao, P. P. (2011). Statistical modelling of outliers for fast visual search. *IEEE VCIDS 2011.*] of SIFT’s scale, orientation, distance to the center, peak strength, …, etc:
  \[
  r(\sigma^*, \theta, D, d, \rho, p_{\sigma\sigma}) = f_1(\sigma^*) \cdot f_2(\theta) \cdot f_3(d) \cdot f_4(D) \cdot f_5(\rho) \cdot f_6(p_{\sigma\sigma})
  \]
- Use self-matching (m29359) to improve the offline repeatability stats robustness

Self-matching via random out of plane rotation
Illustration of FS via offline repeatability PMF

SIFT peak strength pmf

SIFT scale pmf

Combined scale/peak strength pmf
Feature Selection - Repeatability Model

- SIFT repeatability model
Global Descriptor

Why need global descriptor?

- Key points based query representation is not stateless, it has a structure, i.e., SIFTs and their positions. This is not good for retrieval against a large database, complexity $O(N)$
- Need a “coarser” representation of the information contained in the image by aggregating local features, for indexing/hashing purpose.

CDVS Global Aggregation Works

- m28061: Beijing University SCFV: for retrieval/identification
- m31491: Samsung AKULA: for matching/verification
- M31426: Univ of Surrey/VisualAtom: RVD: similar to SCFV
Global Descriptor – SCFV (m28061)

Beijing University SCFV – Scalable Compressed Fisher Vector

- PCA to bring SIFT down from 128 to 32 dimensions
- Train a GMM of 128 components in 32 dim space, with parameters \( \{u_i, \sigma_i, w_i \} \)
- Aggregate \( m=300 \) SIFT with GMM via Fisher Vector, 1\textsuperscript{st}, and 2\textsuperscript{nd} order,

\[
\begin{align*}
g^X_{u_i} &= \frac{\partial \mathcal{L}(X|\lambda)}{\partial u_i} \\
&= \frac{1}{\sqrt{300w_i}} \sum_{t=1}^{300} \gamma_t(i) \left( \frac{x_t - u_i}{\sigma_i} \right)
\end{align*}
\]

where \( \gamma_t(i) \) is the prob of SIFT \( x_t \) being generated by GMM component \( i \),

\[
\begin{align*}
g^X_{\sigma_i} &= \frac{\partial \mathcal{L}(X|\lambda)}{\partial \sigma_i} \\
&= \frac{1}{\sqrt{600w_i}} \sum_{t=1}^{300} \gamma_t(i) \left[ \left( \frac{x_t - u_i}{\sigma_i} \right)^2 - 1 \right] \\
\gamma_t(i) &= p(i|x_t, \lambda) = \frac{w_i p_i(x_t|\lambda)}{\sum_{j=1}^{128} w_j p_j(x_t|\lambda)}
\end{align*}
\]
The SCFV has 32x128 bits, for the 1\textsuperscript{st} order Fisher Vector, and additional 32x128 bits for the 2\textsuperscript{nd} order FV.

Not all GMM components are active, so an 128-bit flag \([b_1, b_2, \ldots, b_{128}]\) is also introduced to indicate if it is active. Rationale: if not many SIFTs are associated with certain component, then the bits it generates are noise most likely.

Distance metric:

\[
s_{X,Y} = \frac{\sum_{i=1}^{128} b_i^X b_i^Y w_{Ha(u_i^X, u_i^Y)} (32 - 2Ha(u_i^X, u_i^Y))}{32 \sqrt{\sum_{i=1}^{128} b_i^X} \sqrt{\sum_{i=1}^{128} b_i^Y}}
\]

A lot of painful work on GMM component turn on logic optimization to reach a very high performance.

Very fast due to binary ops, can short list a 1m image data base within 1 sec on desktop computer.
SCFV Performance

- Matching/Verification (TPR @ 1% FPR)

- Retrieval/Identification (mAP)
SIFT Descriptor Compression

- VisualAtoms/Univ of Surrey: a handcrafted transform/quantization scheme + huffman coding, low memory cost, slightly less performance (compared to PVQ), adopted.
- Only binary form is received, cannot recover the SIFT

\[ h_1, h_2, h_3, h_4, h_5, h_6, h_7 \]

Transform A
\[
\begin{align*}
v_0 &= (h_2 - h_6)/2 \\
v_1 &= (h_3 - h_7)/2 \\
v_2 &= (h_0 - h_1)/2 \\
v_3 &= (h_2 - h_3)/2 \\
v_4 &= (h_4 - h_5)/2 \\
v_5 &= (h_6 - h_7)/2 \\
v_6 &= ((h_0 + h_4) - (h_2 + h_6))/4 \\
v_7 &= ((h_0 + h_2 + h_4 + h_6) - (h_1 + h_3 + h_5 + h_7))/8
\end{align*}
\]

Transform B
\[
\begin{align*}
v_0 &= (h_0 - h_4)/2 \\
v_1 &= (h_1 - h_6)/2 \\
v_2 &= (h_7 - h_0)/2 \\
v_3 &= (h_1 - h_2)/2 \\
v_4 &= (h_3 - h_4)/2 \\
v_5 &= (h_5 - h_6)/2 \\
v_6 &= ((h_1 + h_5) - (h_3 + h_7))/4 \\
v_7 &= ((h_0 + h_1 + h_2 + h_3) - (h_4 + h_5 + h_6 + h_7))/8
\end{align*}
\]

\[ i\tilde{v}_j = \begin{cases} 
-1 & \text{if } i v_j \leq iQL_j \\
0 & \text{if } i v_j > iQL_j \text{ and } i v_j \leq iQH_j \\
+1 & \text{if } i v_j > iQH_j
\end{cases} \]
Outline

- The Problem and CDVS Standardization Scope
- CDVS Query Extraction and Compression Pipeline
  - Key point Detection and Selection (ALP, CABOX, FS)
  - Global Descriptor: Key points aggregation and compression (SCFV, RVD, AKULA)
  - Local Descriptor: Key points and coordinates compression
- CDVS Query Processing
  - Pair-wise Matching
  - Retrieval
  - Indexing
- CDVA Work
  - Handling video input
  - Image/Video Understanding
- Summary
Pair Wise Matching

Diagram of Image Matching

- First local features are matched,
- if certain number of matched SIFT pairs are identified, then a Geometric Verification called DISTRAT is performed, to check the consistence of the matching points via distance ratio check.
- For un-sure image pairs, the global descriptor distance is computed and a threshold is applied to decide match or non-match.
Matching Performance (@ 1% FPR)

- Image Matching Accuracy:
  - Mix of graphics, landmarks, buildings, objects, video clips, and paintings.

- Image Identification Accuracy:
  - For graphics (cd/book cover, logos, papers), paintings, the performance is in 90% range.
  - For objects of mixed variety, 78% in average.
  - For buildings/landmark, the performance is not reflective of the true potential, as the current data set has some annotation errors.
CDVS Retrieval

- **Retrieval Pipeline**
  - Short list is generated by GD based k-NN operation via:
    \[
    d(R, Q) = \frac{\sum_{i=1}^{512} b_i^Q b_i^R W_1 H_a(u_i^Q, u_i^R) W_2 u_i^Q (D - 2H_a(u_i^Q, u_i^R))}{(\sum_{i=1}^{512} b_i^Q)^{0.3} (\sum_{i=1}^{512} b_i^R)^{0.3}}
    \]
  - Then for the short list of m candidates, do m times local descriptor based matching and rank their matching scores
mAP example

- MAP is computed across all query results as the average precision over the recall.

\[
\text{MAP} = \frac{0.564 + 0.623}{2} = 0.594
\]

- MAP favours systems which return relevant documents fast.
- Precision-biased
CDVS Retrieval Performance

- Retrieval Simulation Set Up
  - Approx. 17k annotated images mixed with 1m+ distraction image set
  - Short Listing: retrieve 500 closest matches by GD and then do pair wise matching and ranking
  - Data sets

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Images" /></td>
<td><img src="image2.png" alt="Images" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3.png" alt="Images" /></td>
<td><img src="image4.png" alt="Images" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5. Common Objects</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image5.png" alt="Images" /></td>
<td><img src="Images/bar_graphs.png" alt="Bar Graphs" /></td>
</tr>
</tbody>
</table>

mAP
MBIT (multi-block index table) Indexing

- GD is partitioned into blocks of 16 bits, and inverted list built.
- Shortlisting is by weighted scoring on block wise hamming distance

[Bit Mask/Collision Optimized Indexing]

- Selecting 6-bits segments that are most efficient in discriminating for shortlisting, allow for permutation of bits
- Shortlisting by weighted segment hamming distance also reflecting the segment entropy
Outline

- ReCap of Lecture 10
  - Logistic Regression
  - SVM

- Visual Object Re-Identification
  - MPEG CDVS Research
  - Google Landmark Challenge

- Summary
2018 Google Landmark Recognition Challenge

- Data Set:

- Challenges
  - Recognition: 1:1
  - Retrieval: 1:N
  - Dataset URL: https://www.kaggle.com/google/google-landmarks-dataset
Deep learning features aggregation
- Classification loss not working
  - Instead need to learn a metric:
    - $x = f(I)$: deep learning to extract feature $x$: eg. $512 \times (7 \times 7)$ from VGG16
    - $y = g(x)$: learn a local (linear) metric: e.g. $y = FV(x)$
Random Samples Consensus

- select 4 matching pairs in random, run SVD based homography estimation (HW-1) from it, and check how many points pairs are in agreement
- keep doing until we select the one that best fits
- SoftMax not working (1m labels !)
  - Don’t train a landmark classifier
  - Metric learning for training descriptors or use pre-trained
  - k-NN classifier

- Combine CNNs with classical approaches
  - Global CNN-based descriptors (GeM)
  - Local features and spatial verification (SP class)

\[
C = f(A) + \lambda \cdot g(B)
\]
Rank #3 solution for Google Landmark Challenge

- GeM to create a shortlist:

  ![Diagram](image)

  - Spatial Matching for re-ranking: DELF local feature

    - Tentative matches (ASMK)
    - Spatial verification (SP) [5]

Summary

- MPEG CDVS offers the state-of-art tech performance in visual object re-identification accuracy, speed, and query compression.

- Amd work:
  - A recoverable SIFT compression scheme, currently SIFT cannot be recovered from bit stream.
  - 3D key points, wide adoption of RGB+Depth sensors.
  - Non-rigid body object identification.

- What deep learning brought:
  - Richer set of features from classification networks, e.g. VGG, ResNet.
  - Aggregation opportunity at different stages of convolutions.
Key References

- Test Model 11: ISO/IEC JTC1/SC29/WG11/N14393
- SoDIS: ISO/IEC DIS 15938-13 Information technology — Multimedia content description interface — Part 13: Compact descriptors for visual search
- ALP: m31369 CDVS: Telecom Italia’s response to CE1 – Interest point detection
- RVD: m31426 Improving performance and usability of CDVS TM7 with a Robust Visual Descriptor (RVD)