Image-Based Ordinal Regression of Snow Depth: 
A deep learning approach.

EGU 
ITS4.6/NH6.7 
Data Science and Machine Learning for Natural Hazards and Seismology 

Pierre Lepetit (Météo-France, LATMOS)

Cécile Mallet (LATMOS), Laurent Barthès (LATMOS)
Observation of the snow cover: why and how.

In French plains, snow accumulation is a rare, local and hardly predictable phenomenon which could have heavy consequences on the road transports. Nowadays, monitoring of snow depth mainly relies on human observation, satellite teledetection and local specific sensors.

On the one hand, human observation and local sensors are expensive and teledetection through snowfall is challenging. On the other hand, webcam data became ubiquitous and progress has been made in image-based estimation of meteorological parameters [8, 11, 6, 12, 17, 9, 5].

Here, we present an approach of the **quantitative characterization** of the snow cover based on webcam images. We train a Machine Learning model to yield a quantitative index correlated with the snow depth measured by specific sensors.
Scope of the study

To train and test our model on relevant images, we built our datasets from daytime and nighttime images taken around starts of snow events. We limited the corpus to RGB images taken by road webcams, because they form the main part of the available webcam data, and because they contain critical informations for operational meteorology.

On these images, the difficulties lie in the diversity of the road scenes, the heterogeneity of the devices and in the noises due to adverse weather conditions (see fig.5 - 15% of our images are corrupted). Machine Learning approaches based on wide datasets may yield interesting results while facing these difficulties. But in our case, there is an obstacle: there’s a lack of accurate quantitative labels.
Previous work

On the detection and the quantification of the snow:
Previous studies on image-based snow characterization deal with classification tasks (e.g. snow vs. no snow) [3, 4, 5]. In these studies, the datasets are not conceived to analyze the snow growth during snowfalls. Images are generally taken around midday [3] and by good weather [3,4,5].
Moreover, to our knowledge, there has been no attempt to rank snow depth by machine learning.

On image-based weather estimation:
- End-to-end deep learning approaches perform better, as in other domains of SUN [6,7,9].
- Pairwise learning with siamese networks is efficient for both classification and regression tasks [9,10,13].
- Weak supervision with handcrafted binary comparisons (ordinal labels) mitigates the lack of quantitative labels [10].

We also use ordinal regression to yield a quantitative index. Our approach is based on an end-to-end ranking model. The training phase relies on a directed graph built from handcrafted ordinal labels.
Building the datasets: **three sources** of webcam images

**Collection of AMOS11000:**

AMOS [2] is a large database of worldwide webcam images. To sample images of snow events, geo-stamp and Time-stamp of AMOS webcams have been cross-checked with the height of snow from the ERA-5 reanalysis [20].

A manual subsampling has then been made to avoid redundancy, dark nights and strongly corrupted images.

**Collection of DIR7000:**

The webcam network of the Direction Interdépartementale des Routes was archived during the snow event of the 23-26/01/2019. The time-step is generally finer than for AMOS sequences and the growth of the snow cover is better sampled.

**Collection of TENEBRE-q:**

The webcams of the TENEBRE network, owned by Météo-France, are hosted in weather stations. Their archives have been resampled. All the images associated with snow, bad visibility and rain have been kept. Our TENEBRE-q dataset was then completed by 11000 images taken by good weather (~1000 by scene).
Fig. 9. Sets of webcam sequences used in this study. The blocks illustrate the structure of the data.

**Handcrafted labelling of AMOSDIR18000 (and TENEBRE-h):**

Qualitative labels: 4 levels of snow cover
- 0: « no snow » (20% of the dataset)
- 1: « snow settles on the ground » (40%)
- 2: « snow settles on the road » (30%)
- 3: « ground and road are totally covered » (10%)

Ordinal labels: 10,400 pairs of consecutive images
- Labels « > » or « < »: 6,500 pairs
- Label « = »: 3,900 pairs

**Instrumental labelling of TENEBRE-q:**
Labels come from snow-depth sensors (JENOPTIK SHM30 and APICAL TLN35R) colocalized with the webcams of TENEBRE. Values are given in centimeters with an incertitude of ± 0.5 cm.

**Dual labelling of TENEBRE-h:**
A subset of TENEBRE-q has been manually labelled. These images have hence both instrumental and handcrafted labels. It allowed in particular to check the quality and the coherence of both modes of labelling.
We apply rules to convert ordinal and qualitative labels into partial orders on image sets. These partial orders are stored in four Directed Graphs: $\text{DG}_{\text{train-o}}$, $\text{DG}_{\text{train-q}}$, $\text{DG}_{\text{val}}$ and $\text{DG}_{\text{test}}$.

### Rules for ordinal labels:

1. **Label conversion**:
   - $\text{image } x_1 > \text{image } x_2 \rightarrow \text{new edge } (x_1, x_2)$
   - $\text{image } x_3 < \text{image } x_4 \rightarrow \text{new edge } (x_4, x_3)$
   - $\text{image } x_2 = \text{image } x_4 \rightarrow \text{new edges } (x_1, x_4), (x_3, x_2)$

2. **Take the transitive closure**

### Rules for qualitative labels:

1. **Level conversion**:
   - Level 0 < levels 1; 2; and 3. Level 1 < level 3.

   - **Images at level 0**: « no snow »
   - **Images at level 1**: « snow on ground »
   - **Images at level 2**: « snow on road »
   - **Images at level 3**: « white road »

2. **Keep randomly 100 edges max. by webcam sequence**

**AMOSDIR18000** is split into training webcam sequences (80%) and validation webcam sequences (20%). $\text{DG}_{\text{train-o}}$, and $\text{DG}_{\text{train-q}}$, are made with labelled images of the train sequences. $\text{DG}_{\text{val}}$ is made with the validation sequences. $\text{DG}_{\text{test}}$ is made with images of **TENEBRE-h**.
Training along the graphs

**Edge selection:**
- At each step, edges are selected:
  - from $D_{\text{train-o}}$ with a proba. of $p_o$
  - from $D_{\text{train-q}}$ with a proba. of $p_q = 1 - p_o$
- Inside $D_{\text{train-o}}$, a weighted sampling is used to favour edges between small sequences.

**Models:**
Two kinds of architectures were tested: Resnet50 [14] and VGG16 [15]. The softmax layer is replaced by a one-dimensional output layer.

**Loss Function:**
- Built on Ranking Hinge Losses
- The first member penalizes miss-ordered outputs.
- The other terms force the output positivity.

**Other Training parameters:** Classical data-augmentation tools were applied. Mini-batch count 64 cropped images. The gradient descent is optimized through the ADAM algorithm [23] with a starting learning rate of 0.001.
## Results on ranking tasks and model selection

<table>
<thead>
<tr>
<th>Pretrained models</th>
<th>VGG16 (places365)</th>
<th>VGG16 (imagenet)</th>
<th>Resnet50 (places365)</th>
<th>Resnet50 (imagenet)</th>
</tr>
</thead>
<tbody>
<tr>
<td>validation (best accuracy)</td>
<td>83.6 %</td>
<td>82.1 %</td>
<td>87.3 %</td>
<td>88.3 %</td>
</tr>
<tr>
<td>test (accuracy)</td>
<td>69.9 %</td>
<td>73.3 %</td>
<td>74.2 %</td>
<td>74.8 %</td>
</tr>
</tbody>
</table>

Tab.1. Performances of pretrained CNN

During the validation and the test phases, all the edges of $D_{val}$ and $D_{test}$ are browsed. The accuracy is the fraction of the output pairs that are correctly ordered. We stress the fact that the webcams of the validation and test sequences were not used during the training phase.

To evaluate the correspondence with the instrumental labels, we will use the Resnet50 pretrained on Imagenet with $p_o = 0.7$, $p_q = 0.3$ and $a = 0.1$ (hinge margin). Its output is now referred to as our snow depth index. As we still meet problems with nighttime images, the following results only hold for the daytime part of the TENE BRE-q dataset.
Evaluation on TENEBRE-q (daytime images)

Two dimensional histograms show the extent to which our snow depth index co-vary with the instrumental labels. On these examples, the two first behave correctly. For both Roissy and Dorans1 webcams, the dispersion is stronger, and for the four lasts, predictions is more hazardous. Whatever the reasons (changes in the webcam orientation, instrumental default, etc), the main point is that the range of our snow depth index seems to be relative to the view. We are still working on that issue.

Finally, the rank correlations are lower over the 5cm – 15 cm interval. We ignore if it comes from the scarcity of such snow depths in our training dataset.
Evaluation on TENEBRE-q (daytime images):

Fig. 12. Coevolution of our snow depth index (after rescaling) and the instrumental labels. Strong discontinuities are partly due to concatenation of daytimes. The same affine rescaling has been used for all the curves.

Time series allow to observe the same phenomenon: if snow periods are well delimited by the index, the index range varies from one webcam to the other. Moreover, the prediction is affected by a high-frequency noise which may be inherent to our methodology or due to quick variations of correlated environmental factors, as the illumination.
Conclusion and perspectives

- We presented an end-to-end learning to rank framework that yield a snow depth index. On an independant subset of webcams located in weather stations, this index varies coherently with instrumental measures, at least by day, and for the first centimeters. The same approach gives similar results on the problem of the optical range estimation.

- However, the range of the output index varies from one webcam to another and from daytime to nighttime (not shown). It limits the interest for an application in road meteorology. Until now, our naive attempts to combine regression and ranking, in order to calibrate the estimations and improve the ranking correlation, led to overfitting. Nevertheless, customization of the training parameters could allow to benefit from both kinds of label, as suggested in [21].

- Finally, the time series of our snow index display a high-frequency noise that may trigger false detections. We ignore if this noise is inherent to our ranking method or if a better disentangling of correlated factors would help. In the second case, an extension to a semi-supervised framework with a constraint on time-variations may help (see [16]).