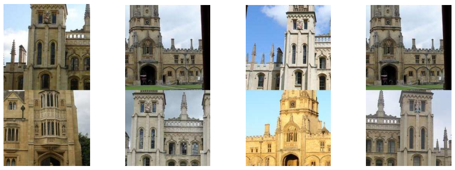

Figure 4: The four most similar false positive patches over the Oxford5k using the CNN-representation.

6.3. Results

The result of four different retrieval methods applied to 5 datasets and the concatenation of the Oxford5k, Paris6k and Sculpture6k are presented in table 7. Figure 4 shows the most similar false positive for the Oxford5k dataset. This gives an indication of why retrieval task for similar buildings can be difficult with a generic (not tuned) representation.

We combined Oxford5k, Paris6k and Sculp6k and applied all 180 queries. The CNN representation achieved a mAP of 60.01 while VLAD reached a mAP of 51.88 percent. VLAD's performance falls below that of the CNN-representation in this test mainly because VLAD does not work for smooth item retrieval. A small technical detail is $L2$ normalization of each CNN feature dimension increases the mAP by approximately 1 percent over the datasets Oxford5k and Paris6k.

The CNN representation performs relatively poorly on the UKbench datatbase as this dataset addresses the challenge of viewpoint change and CNN features are not invariant to in-plane rotations not present in the training data.

Spatial search increase the processing time and memory consumption by $O(h_q^3 \times h_r^3)$. Where h_q and h_r are the height of spatial pyramids over query and reference images respectively. Therefore, the height of the pyramid should be kept as small as possible. The query items in the UKbench and the Holidays datasets are centered and spatial search is not required in either query or reference image. On the other hand the items in the Oxford5k, Paris6k and Sculp6k reference images are not centered and searching through the reference images increases the performance of retrieval.

The difference between landmarks in the Oxford5k and Paris6k datasets are subtle. For example nuances of the window architecture are the most visual distinctive features for many buildings. Hence, intuitively spatial search over the query images also should increase the performance of retrieval over the aforementioned datasets. While the important structures in the Sculpture6k dataset are the global shape of the sculpture and small patches in a query image do not capture this information. Figure 5 shows the effect of spatial search over the Paris6k and Sculpture6k datasets.

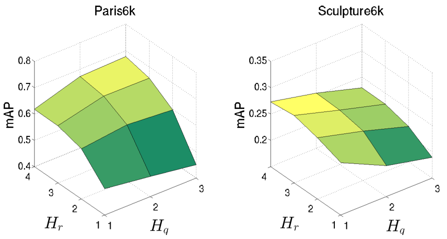

Figure 5: The effect of the spatial pyramid search for Paris6k and Sculp6k. As items in the reference sets are not centered, increasing the pyramid height for reference images constantly increases the performance. In Sculp6k complete shape of the sculpture matters. Therefore, spatial search in the query images decreases the performance. In Paris6k, a nuanced detail like the shape of a window in a building might be all the information we seek, hence increasing the height of pyramid increases the performance of retrieval.

	Oxford5k	Paris6k	Sculp6k	Holidays	Comb.	UKBench
VLAD 64D[4]	**0.555**[4]	0.642	-	**0.646**[4]	0.519	**3.38**
BoW 200kD	0.364[18]	0.460[33]	0.086[3]	0.540[4]	-	2.81[18]
IFV 64D[31]	0.418[4]	-	-	0.626[4]	-	3.35[18]
BoB	N/A	N/A	0.253[3]	N/A	-	N/A
CNN	0.520	**0.676**	**0.269**	**0.646**	**0.600**	3.05

Table 7: **The result of object retrieval on 6 datasets.** All the methods except the CNN have their representation trained on same dataset that they report the results on. The result of VLAD[4] on Oxford5k with a dictionary trained on Flicker60k is 0.478[4]. ”Comb” has the results for the first three datasets combined.

7. Conclusion

In this work, we used an off-the-shelf CNN representation, `OverFeat` , with simple classifiers to address different recognition tasks. The learned CNN model was originally optimized for the task of object classification in ILSVRC 2013 dataset. Nevertheless, it showed to be a strong competitor for the more sophisticated and highly tuned state-of-the-art methods. The same trend was observed for various recognition tasks and different datasets which highlights the effectiveness and generality of the learned representations. The experiments confirm and extends the results reported in [11]. It can be concluded that from now on, deep learning with CNN has to be considered as the primary candidate in essentially any visual recognition task.

References

[1] Imagenet large scale visual recognition challenge 2013 (ilsvrc2013). http://www.image-net.org/challenges/LSVRC/2013/.

[2] A. Angelova and S. Zhu. Efficient object detection and segmentation for fine-grained recognition. In *CVPR*, 2013.

[3] R. Arandjelović and A. Zisserman. Smooth object retrieval using a bag of boundaries. In *ICCV*, 2011.

[4] R. Arandjelović and A. Zisserman. All about VLAD. In *CVPR*, 2013.