Change Detection from a Street Image Pair using CNN Features and Superpixel Segmentation

Ken Sakurada
sakurada@vision.is.tohoku.ac.jp
Tohoku Univeristy
Miyagi, Japan

Takayuki Okatani
okatani@vision.is.tohoku.ac.jp
Tohoku University/JST CREST
Miyagi, Japan

This paper proposes a method for detecting changes of a scene using a pair of its vehicular, omnidirectional images. The top images of figure 1 show an example of such image pairs taken at different times. Apparently, there are temporal differences in illumination and photographing conditions. Moreover, there has to exist visual difference in camera viewpoints, although they were captured from a vehicle running on the same street and were matched using GPS data. This is due to differences in vehicle paths and shutter timing. The type of scene changes targeted here includes 3D (e.g. vanishing/emergence of buildings, cars etc.) as well as 2D changes (e.g. changes of textures on building walls). To precisely detect these changes from such an image pair, it is necessary to overcome these unwanted visual differences.

We tackle the change detection problem in the 2D domain. That is, we consider detecting changes based on the direct comparison of a pair of images. The major issue is then how to deal with the unwanted visual differences (i.e., viewpoint differences etc.) To cope with this, we propose to use the features extracted by convolutional neural networks (CNNs). To be specific, we use a fully trained CNN for large-scale object recognition task [4] in a transfer learning setting. It was reported in the literature that using activation of the upper layers of a CNN trained for a specific task can be reused for other visual classification tasks. Several recent researches imply that the upper layers of CNNs represent and encode highly-abstract information about the input image [1, 2, 5]. We conjecture that highly-abstract (or object-level) changes can be detected by using the upper layers, whereas low-level visual changes (e.g. edge, texture etc.) will be detected using the lower layers. We show that this conjecture is true through several experimental results.

The proposed method consists of the three components: i) extraction of grid features, ii) superpixel segmentation, and iii) estimation of sky and ground areas by Geometric Context. These are described below.

(i) Extraction of grid features We denote two input images by I^t and $I^{t'}$, where t and t' are the times at which they were captured. First, I^t and $I^{t'}$ are divided into grid cells $g(= 1,...,N_g)$. A feature is extracted from each grid cell g, yielding $\mathbf{x}_g{}^t$ and $\mathbf{x}_g{}^{t'}$.

The changes that we want to detect are object-level changes (e.g, the emergence/vanishing of buildings and cars) and not low-level, appearance changes due to changes in viewpoints, illumination or photographing conditions. To distinguish these two, the proposed method uses the activation of a upper layer of a deep CNN for the grid features $\mathbf{x}_g{}^t$ and $\mathbf{x}_g{}^{t'}$. To be specific, we use a pooling layer of the CNN. Each feature (e.g., $\mathbf{x}_g{}^t$) is the activation of all the units in the same location across the maps of the pooling layer. Thus $\mathbf{x}_g{}^t$ has the same number of elements as the maps of the pooling layer.

Next, these features are normalized so that $|\mathbf{x}_g{}^t| = 1$, and then their dissimilarity is calculated at each grid cell g as

$$d_g = |\mathbf{x}_g{}^t - \mathbf{x}_g{}^{t'}|. \tag{1}$$

Then, the dissimilarity d_g is projected onto the input images I^t and $I^{t'}$, determining the pixel-level dissimilarity $d_p(p = 1,...,N_p)$; N_p is the number of pixels. This is done by simply setting $d_p = d_g$ for any pixel p contained in the grid cell g.

(ii) Superpixel segmentation The difference in viewpoint is arguably the major source of difficulties for 2D change detection methods. The use of the CNN features is expected to help mitigate this difficulty, owing to the property of the CNN features invariant to geometric transformation such as translation, 3D rotation, and even more complicated ones. However, the resolution of the dissimilarity map d_p's is basically very low. We use superpixel segmentation to refine the dissimilarity map to hope for obtaining precise boundaries of the detected changes.

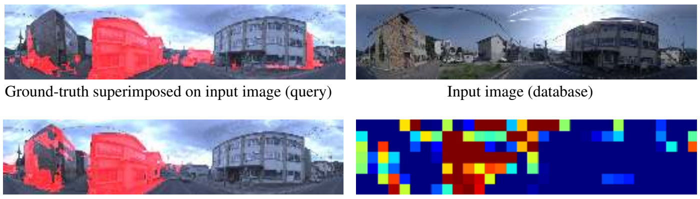
Ground-truth superimposed on input image (query)
Input image (database)
Change estimation (binarized)
Feature distance (each grid)

Figure 1: Results of change detection using pool-5 feature of CNN.

This starts with computing superpixel segmentation of I^t and $I^{t'}$. Let s^t be a superpixel and S^t be the set of superpixels. We define the dissimilarity d_{s^t} at a superpixel $s^t \in S^t$ to be the average of all the pixels in s^t as

$$d_{s^t} = \frac{1}{|s^t|} \sum_{p \in s^t} d_p. \tag{2}$$

We denote the maximum value of d_{s^t} and $d_{s^{t'}}$ by $d_{\max}$, i.e., $d_{\max} = \max(d_{s^t}, d_{s^{t'}})$.

(iii) Estimation of sky and ground areas by Geometric Context In the last step of the proposed method, Geometric Context [3] is used to remove the segments of sky and ground from the images. Geometric Context is a segmentation method that is known to be robust to changes in illumination and photographing conditions. It estimates probabilities of the sky and the ground at each pixel $(p_{\text{sky}}{}^t, p_{\text{ground}}{}^t)$ in the input image I^t. Using these, we remove these areas from the images, converting the dissimilarity at each pixel into $\overline{d_p}$ as

$$\overline{d_p} = \begin{cases} 0 & (((p_{\text{sky}}{}^t > a) \wedge (p_{\text{sky}}{}^{t'} > a)) \\ & \vee((p_{\text{ground}}{}^t > b) \wedge (p_{\text{ground}}{}^{t'} > b))), \\ d_{\max} & (\text{otherwise}) \end{cases} \tag{3}$$

where $a = t_{\text{sky}}$ and $b = t_{\text{ground}}$ are constant values within the range of $0 \leq t_{\text{sky}}, t_{\text{ground}} \leq 1$.

We have created *Panoramic Change Detection Dataset* and used it for the experiments. The data used in this study (the pairs of the omnidirectional panoramic images taken at different time points and the hand-labeled ground-truth of change detection) can be downloaded from our website. The dataset consists of two subsets, named "TSUNAMI" and "GSV." "TSUNAMI" consists of one hundred panoramic image pairs of scenes in tsunami-damaged areas of Japan. "GSV" consists of fifty panoramic image pairs of Google Street View. The size of these images is 224 × 1024 pixels.

Figure 1 shows an example of the results of change detection. It is observed from them that the proposed method was able to correctly detect the scene changes, for example, demolished and new buildings and cars. In some cases, Geometric Context could not accurately estimate sky due to electrical wire and pole, or could not distinguish between the ground and low height object (e.g., debris and car). In contrast, the proposed method was able to accurately detect object-level scene changes.

[1] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *ECCV*, pages 584–599. Springer, 2014.

[2] Yoshua Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[3] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Geometric context from a single image. In *ICCV*, pages 654–661, 2005.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[5] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.