Skin Cancer Detection and Tracking
Using Data Synthesis and Deep Learning

Yunzhu Li, Andre Esteva, Brett Kuprel, Rob Novoa, Justin Ko, Sebastian Thrun
1Peking University
leo.liyunzhu@pku.edu.cn
2Stanford University
esteva@cs.stanford.edu, {kuprel, rnovoa, jmko, thrun}@stanford.edu

Abstract
Dense object detection and temporal tracking are needed across applications domains ranging from people-tracking to analysis of satellite imagery over time. The detection and tracking of malignant skin cancers and benign moles poses a particularly challenging problem due to the general uniformity of large skin patches, the fact that skin lesions vary little in their appearance, and the relatively small amount of data available. Here we introduce a novel data synthesis technique that merges images of individual skin lesions with full-body images and heavily augments them to generate significant amounts of data. We build a convolutional neural network (CNN) based system, trained on this synthetic data, and demonstrate superior performance to traditional detection and tracking techniques. Additionally, we compare our system to humans trained with simple criteria. Our system is intended for potential clinical use to augment the capabilities of healthcare providers. While domain-specific, we believe the methods invoked in this work will be useful in applying CNNs across domains that suffer from limited data availability.

Introduction
Dermatology is a medical field which stands to be heavily augmented by the use of artificial intelligence techniques. Diseases are visually screened for, and many disease diagnoses are performed strictly with an in-clinic visual examination. Discerning between skin lesions is a difficult task - the difference between skin cancer (melanoma, carcinoma) and benign lesions (nevi, seborrheic keratosis) is minute, and the differences are in slight details between them (Figure 1). With 5.4 million cases of skin cancer diagnosed each year in the United States alone, the need for quick and effective clinical screenings is rising (Rogers et al. 2015). Patients with skin cancer tend to be afflicted with many moles, and so one of the challenges in skin cancer screenings is identifying them amongst a myriad of benign lesions. Another key element of these diagnoses is based on inspecting temporal changes in lesions - a fast changing lesion is more likely to be malignant. As such, patients and providers need tools to support this at scale.

Recent advances in detection and tracking using CNNs (Ren et al. 2015; Kanazawa, Jacobs, and Chandraker 2016; Zhou et al. 2016) has the potential to augment healthcare providers by (1) detecting points of malignancy, and (2) finding corresponding lesions across images, allowing them to be tracked temporally. However, the primary challenge in using traditional detection techniques is working in a low-data regime without the availability of high volumes of annotated and labeled data - the largest existing open-source skin cancer dataset of photographic images is the Edinburgh Dermofit dataset, containing 1,300 biopsied images.

To overcome the challenge of working in this low-data regime we develop a domain-specific data synthesis technique which stitches small single-lesion images onto large body images. Both the body images and the skin lesions images are heavily augmented with various techniques, and the lesions are blended onto the bodies using Poisson image editing (Pérez, Gangnet, and Blake 2003).

Figure 1: Key factors for skin cancer care include early detection and tracking over time. Top Row: superficial spreading melanoma, evolving in time (Salerni et al. 2012). Bottom left: comparison between malignant and benign lesions shows the difficulty in early detection. Bottom right: examples of patients afflicted with many lesions.
Figure 2: Data Synthesis. Skin lesions are blended with raw body images to generate detection and tracking data. (Left) Example biopsied skin lesion and raw body images. Top diagrams show the lesion segmentation mask and the gradient field along with semantic regions used to calculate blending locations. (Middle) Generated training images for detection and corresponding label masks. Red areas represent blended malignant lesions, yellow areas represent blended benign lesions. (Right) Generated training images for tracking, along with a few example pixel-wise correspondences.

As an additional baseline we include a simple comparison between our model and two humans trained on the 1,300 images of the Edinburgh dataset. Our method demonstrates a working end-to-end CNN system capable of tackling two critical diagnostic tasks with superior performance to algorithmic baseline techniques. It is trained with very little original data and thus the techniques demonstrated here can be easily transferred to other data-limited domains.

Data Synthesis
We create a domain-specific data augmentation technique for generating synthetic images from two low-data sources: high-quality lesion images (1,300 biopsy-proven cancers and moles), and body images (400 back, leg, and chest images) whose skin regions have been manually segmented. We first generate images for detection and then further augment them for tracking. These images are intended to mimic the real-world clinical case of patients exhibiting many lesions, some possibly malignant, with the need to track them over time.

Detection data is generated in two steps: (1) A blending position on the body image is chosen using local feature matching between a lesion image and the body image, and (2) the lesion image is blended into the body image using Poisson image editing (Pérez, Gangnet, and Blake 2003).

For tracking, data is generated by further augmenting detection images - that is, given a detection image, we create a pair of images from it. The purpose here is to recreate temporal images (which will exhibit changes in lesion shape/size, as well as body changes) and to force the network to learn pixel-wise correspondence by focusing on the distortion in local texture information.

System Pipeline
Our system is composed of two parts: the first detects malignant and benign skin lesions, the second tracks them across images. The detection network is trained with the synthetic images (skin lesion + body images) described in the previous section, using pixel-wise labels. Once the detection network is trained to convergence, its weights are used to initialize the tracking network. This network is then trained on image pairs formed from the detection data. It is worth noting that this particular choice of architecture is largely the result of iterating on potential architectures to determine the one with superior performance.

Detection System
The detection component is intended to highlight to a clinician the potentially malignant lesions on a given input image. Our system feed-forwards an input image through the CNN, outputs a pixel-wise heat-map over the five classes of interest, and then uses post-processing techniques to generate the region proposals.

The network structure, shown in Figure 3, is composed of a convolutional component and a deconvolutional component with skip-link connections as described in (Ronneberger, Fischer, and Brox 2015). These component networks have proven to be effective at pixel-wise predictions and biomedical segmentation (Long, Shelhamer, and Darrell 2015). Examples of raw prediction results and post-processed images are shown in Figure 5.

Tracking System
The tracking component of our system is intended to find pixel-wise correspondence between two images of the same body part, in order to track lesions over time.

Shown in Figure 3, the tracking network is an adaptation of the detection network. Tracking data comes in image pairs - during the feedforward pass the two images are fed through the conv-atrous pipeline independently (Agrawal, Carreira, and Malik 2015; Chen et al. 2014), after which their output is element-wise subtracted before being fed through subsequent convolutional layers (Kanazawa, Jacobs, and Chandraker 2016) to output a 2D vector field of correspondences from the first to the second image of the pair. As a final step, our system searches for the best feature match within a square region of the predicted correspondence point in order to calculate the final correspondence point.
Experiments and Results

We use 1,300 biopsied skin lesion images and 400 high-resolution body images to generate 40,000 images for detection and 84,000 pairs of images for tracking.

Detection

Example results for detection are shown in the first three rows of Figure 5, whose columns show examples as they are processed from input images to raw network output to post-processed results.

The baseline classifier (bottom row of Figure 5) is Google’s Inception-V3 network architecture (Szegedy et al. 2015) pretrained on ImageNet (Russakovsky et al. 2015) and finetuned on the 1,300 skin lesions of our dataset. We extract patches at various scales from the test image using a sliding window technique over the entire image, then collect the label maps and aggregate them together into the 3-way heatmap. Afterwards, we apply the same post processing step that we use for our system in order to generate the region proposals. For comparison, we include the performance of two non-expert humans trained to detect malignant lesions and tested on the same 108 images as our detection CNN. ROC curves are shown in the left part of Figure 4.

Tracking

For tracking, we compare our results to two baseline techniques: SIFT Flow, and Deformable Spatial Pyramids (Liu et al. 2008; Kim et al. 2013).

We evaluate tracking accuracy using the percentage-of-correct-keypoints (PCK) metric. Our test set is composed of 260 pairs of correspondence labels manually annotated on temporal image pairs. These image pairs vary in pose, background, distance, viewpoint, and illumination condition. Example image results are shown in Figure 5, where green lines show correctly predicted correspondences, and red lines show incorrect predictions. The PCK of our method, both with and without feature matching incorporated, is plotted in Figure 4, in comparison to both baselines.

Conclusion

Here we show large-scale detection and tracking of skin lesions across images using FCN in a low-data regime using domain-specific data augmentation. In the absence of large amounts of labeled and annotated data, we generate high volumes of synthetic data using 1,300 biopsy-proven clinical images of skin lesions and 400 body images. Skin lesion images are blended onto body images, heavily augmented with a variety of techniques, and used to train a
detection network. We demonstrate human-interpretable detection with this method, and demonstrate superior performance over baseline. We then further augment the data and generate image pairs with pixel-wise correspondence between them, and use this to train a tracking network whose architecture is partially composed of the detection network and initialized with its trained weights, outperforming both SIFTFlow and DSP. The networks are trained on synthetic data and tested on real-world data.

AI systems of this sort have the potential to improve the way healthcare is practiced, which may extend outside of the clinic. Algorithms such as these could aid providers at spotting suspicious lesions amongst benign ones, and at observing temporal changes in lesions that may signify malignancies.

References

Agrawal, P.; Carreira, J.; and Malik, J. 2015. Learning to see by moving.


Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. CVPR.


