SEMI-AUTOMATIC LABELLING OF AERIAL VIDEOS VIA LABEL TRANSFER

Md. Shahid, Abhishek B., N. Y. Reddy, V. R. Khaja, Sumohana S. Channappayya

Indian Institute of Technology Hyderabad

ABSTRACT

Labelled data are of utmost importance for supervised learning in general, and for deep learning in particular. This has resulted in the creation of large scale labeled image and video data sets to address computer vision problems using deep learning models. The ImageNet data set is a classic example of such a data set. While several such labelled data sets have been created for images of objects, faces, animals, activities, street views etc., there are few that contain aerial imagery (and corresponding labels). We address this shortcoming by proposing a semi-automatic labelling strategy for aerial imagery using label transfer. Specifically, we generate aerial imagery (frames from videos) using the Google Earth Engine (GEE). We are releasing typical dataset file. These frames are manually labelled (semantically segmented) at periodic intervals and the labels for the intermediate frames are predicted using a simple algorithm. We use sparse key point detection over these frames to generate homography parameters. These transformation parameters are applied to the manually generated labels and merged appropriately to generate the predicted labels. Our proposed approach clearly outperforms existing approaches on our GEE data set both qualitatively and quantitatively.

1 Introduction

The availability of large volumes of data combined with the exponential growth of computational resources heralded the deep learning era. However, labeled data is crucial for learning effective representations using deep learning models. Generating manually labelled data is a very time consuming and expensive process. This has led to work on label transfer approaches that help alleviate some of the issues associated with the manual process. Feature matching is a key ingredient to label transfer. The literature is rich with many feature matching approaches like SIFTflow [1], DSPflow [2], ProposalFlow [3], CNNflow [4], and some of them have been extended for dense label transfer. For e.g., Zhu et al. [5] proposed semantic segmentation using a pre-trained network (DeeplabV3+ [6]) for video prediction, reconstruction and joint propagation of labels. They demonstrated an improvement in mean Intersection over Union (mIoU) of video reconstruction with joint label propagation over video prediction and just label propagation.

While there are several labelled data sets of terrestrial images taken at the ground level, there are few publicly available labelled data sets with aerial imagery. The INRIA data set [7] is one of the few labelled aerial image data sets. However, it has only two labels (*building*, *non-building*) that limits its utility. This lack of labelled aerial image data sets makes the training of deep learning models difficult and leads to poor generalization on such images. For e.g., a state-of-the-art object detect algorithm like YOLO v3 [8] does not perform well on aerial image data. Further, the justification for creating labelled aerial image data sets comes from the ever-increasing use of drones and unmanned aerial vehicles for a variety of applications ranging from surveillance to disaster management to emergency response etc.

To address these requirements, we propose a label transfer based semi-automatic approach for labelling aerial imagery acquired using the Google Earth Engine (GEE). The proposed method facilitates the generation of labelled aerial image data sets that can then be used for training deep learning models to solve computer vision problems such as semantic segmentation, cross-domain matching, object detection and tracking etc.

2 Background

We present a brief review of the INRIA aerial image data set[7] and label transfer techniques to facilitate further development.

2.1 The INRIA Data Set

The INRIA aerial image labeling benchmark [7] is one of the few labelled data sets containing aerial images with two classes {*building*, *non-building*}. This work demonstrated generalisation by training classifiers over images from certain cities and testing over other cities, and concludes that neural networks generalise reasonably well. To further evaluate the generalisability of these models, we trained a fully convolutional network (FCN) based classifier over the training data in the INRIA data set and tested it over aerial images of a city in India. The performance of this model is shown in Fig. 1. The first row of this figure shows the output of the classifier on an image from the INRIA dataset and the second row shows the output on an image from our data set. Clearly, the model does not generalise well on our data set (of aerial images from a

city in India). This essentially means that the classifier needs to be trained again on our data set. The challenge however is that we do not have the ground truth labelled data for our aerial imagery.

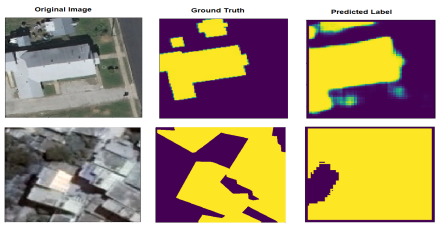

Fig. 1: Prediction of a trained FCN based classifier over INRIA test image (top row) and our data set (bottom row).

2.2 Label Transfer Methods

We briefly review conventional approaches for transferring labels over partially labelled data sets (of videos). Typically, these methods rely on finding matching features (correspondence) either in the spatial or in the temporal domain. The homography associated with the feature correspondence is then applied to the manual labels to find the predicted labels. Traditional feature correspondence approaches include those based on optical flow [9], scale invariant feature matching such as SIFT [10] etc. SIFTflow [1] consists of matching densely sampled, pixel-wise SIFT features between two images, while preserving spatial discontinuities. It preserves the spatial model and allows matching of objects located at different parts. It builds correspondence of SIFT features pixel-wise instead of RGB gradient that was used in optical flow. The same is used for label propagation. It uses learned motion vectors over optical flow (CNN flow estimator-FlowNet2 [11]). It is primarily intended for autonomous navigation applications with propagation length limited to ± 5 frames.

DSPflow [2] uses a deformable spatial pyramid (DSP) matching algorithm for computing dense pixel correspondences. Dense matching involves appearance between pixels and geometric smoothness between neighboring pixels. A pyramid graph model that regularizes match consistency at multiple spatial extents is used to handle "deformable" regions as opposed to the strict rigidity of traditional spatial pyramids.

Unlike semantic flow approaches used in SIFTflow and DSPflow, Proposalflow [3] exploits modern object proposals, that exhibit high repeatability at multiple scales, and can take advantage of both local and geometric consistency constraints among proposals. Proposal flow boosts with learning-based descriptors for semantic correspondences, or learning geometric matching.

Deep learning methods have significantly outperformed traditional statistical and machine learning methods in the recent past on various imaging and non-imaging tasks. We briefly summarize the performance of three recent deep learning based feature matching/object detection/image segmentation methods on aerial imagery for an Indian city.YoloV3 [8] is a state-of-the-art object detection algorithm that we apply on our data set to detect common object classes. These results are shown in the left column of Fig. 2.We can see it detects one or two cars correctly out of several cars on the road and misclassifies many other classes (boat, person etc.). SegNet [12] is a

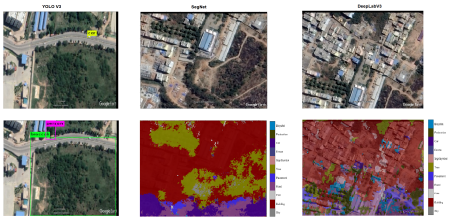

Fig. 2: Deep learning based object detection, and segmentation methods evaluated on example aerial images from our data set.

deep encoder-decoder architecture for multi-class pixel-wise segmentation with VGG16 model. It has been designed and trained for urban road scene segmentation. It consists of 12 classes including roads, trees etc. The segmentation results of the algorithm on our data set is shown in the middle column of Fig. 2.

DeepLab [13] is a state-of-art deep learning model for semantic image segmentation, where the goal is to assign semantic labels to every pixel in the input image. DeepLabv3+ [6] employs the encoder-decoder structure where DeepLabv3 is used to encode the rich contextual information. It uses resnet18 model trained over camvid data set [14]. The DeepLabv3+ based semantically segmented aerial image is shown in the right column of Fig. 2 (right column). From these results, it is clear that existing state-of-the-art object detection and segmentation methods cannot be directly applied to aerial images. These observations provide the primary motivation for this work. Specifically, we provide a low complexity semi-automated labelling approach to generate labels of aerial imagery. Also, we provide fine grained labels (compared to the INRIA data set[7]) that allows for applications such as semantic segmentation.

3 Proposed Approach

We first describe our method for acquiring aerial images using the Google Earth Engine (GEE) [15] and then describe the proposed semi-automatic label transfer method.

3.1 Data Set Generation

We have collected airborne data by manned aircraft in a designated urban area. Video data is stored on-board and trajectory data is transmitted via RF link and stored in a

ground control station. There is no synchronisation between stored video (DTV) and telemetry/flight data. These parameters data are noisy due to RF communication(LOS disturbances/environmental conditions). We have extracted airborne video data from the publicly available GEE [15] for the same trajectory. Our choice was driven by the observation that the INRIA [7] data set focused on cities in the United States and Europe. We chose an Indian urban locality to not only provide for variety but also to test for generalisability of computer vision algorithms. We selected our trajectory such that it covers various urban regions like roads, buildings, vegetation etc. We acquired this data over varying time scales - months, days, different times of the day and so on. Our goal was to best simulate airborne flight with camera-mounted-at-belly scenario in forward direction. Blue and red buildings are typical landmarks in the data set. A typical kml file can be downloaded from our lab website. We have manually labelled the data into 20+ classes including building, road, mud road, street, runway, vehicle, dry field, vegetation field, construction site, prominent buildings, red/blue building, runway, water and so on. Regions that do not fall into commonly found urban scenery are labeled with a unique label (e.g., 255). A few video frames and corresponding manual labels from our video data set are shown in Fig. 3.

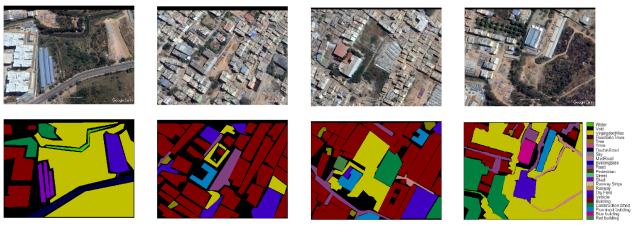

Fig. 3: Samples from the labelled video data set.

3.2 Semi-automatic Label Transfer

In the proposed approach, we first label images manually at a regular intervals with the interval depending on scene movement (content variation over the video). Labels for in-between frames are predicted (transferred) automatically, thus making our approach semi-automatic. Specifically, we propose a two-pass approach for label transfer composed of a forward pass and a backward pass. Our approach is based on the observation that aerial video frames are highly correlated over reasonably large temporal neighborhoods (depending on aircraft speed, altitude, camera zoom, view angle etc.) We empirically observed that major scene changes happen only after a fairly large number of frames (for e.g., 50-100 frames in a video shot using a satellite at 30 frames per second). By forward labelling we mean that the manual label of the current frame is transfer to the subsequent frame (temporally). By the same token, backward labelling refers to the transfer of the manual label to the previous frame in time. The frame with the manual label is compared with subsequent/later frame (corresponding to forward/backward pass) using SURF [16] key point match. Outliers are removed from these points using RANSAC [17]. Geometrically transformed parameters are generated using points after excluding outliers. These parameters are applied to transform the manual label to the predicted (transferred) label. Forward labeling approach is shown in Fig. 4

During forward pass of label transfer in subsequent frames, only labels with correspondence are transferred and the remaining are denoted with zeros. Similarly, during the backward pass, only labels with correspondence are transferred. Forward and backward passed labels are merged with the priority being given to major frames. Forward/backward transferred

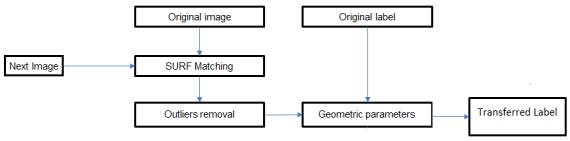

Fig. 4: Label transfer methodology.

labels for contemporary methods are shown in Fig. 5. The first and third row represent forward transfer label for the left original image in same row for contemporary and proposed methods. Similarly, second and fourth row represent backward transfer label. Merging strategy The labels transferred (predicted) using the forward and backward methods are merged appropriately to create a single label for the intermediate video frame. The geometric transformation process leads to artifacts near region boundaries. These label artifacts are corrected by doing a closest label detection. Among these corrected forward and backward labels, major (closer manual label) and minor frames (further manual label) are identified based on geometric parameters derived during label transformation. We noticed a few boundary artifacts due to the merging process. The overall merging approach is shown in Fig. 6.

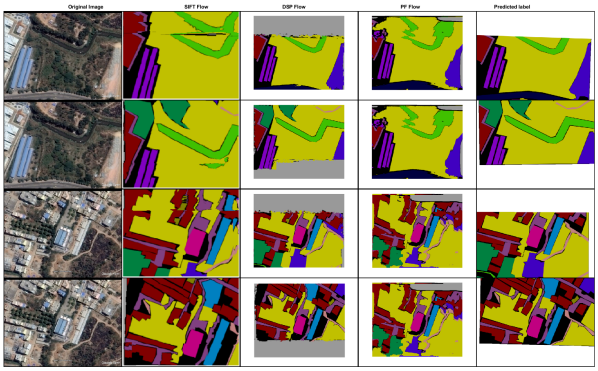

Fig. 5: Forward and backward transferred label

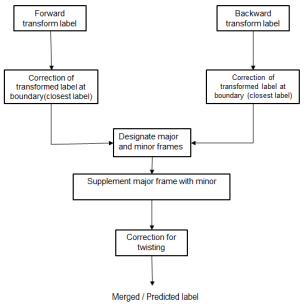

Fig. 6: Merging strategy.

4 Results and Discussion

To evaluate the performance of the proposed and existing methods, we manually labelled intermediate images (e.g., frame number 25, 75, 125 and so on) in addition to labelling frame numbers 1, 50, 100 and so on. The images with frame index 25, 75 etc. are used for testing. The merged label for a given test image are compared with manually labelled images as shown in Fig. 7. For state-of-the-art feature matching approaches, the geometric transformation for label transfer is performed using the matching points given by these methods. Forward and backward labels for state-of-the-art methods are merged manually to ensure fairness. Mean IoU and accuracy [4] are used for quantitative comparison.We have adopted a strategy that intersection is one only if both labels (for each corresponding pixels in ground truth and predicted label are equal) and zero otherwise. Our implementation is done in Matlab on a i7 processor PC. The proposed method requires key-point detector with matching module, followed by geometric transformation. For state-of-the-art methods, we used code available online. Our approach is on par or better with existing methodologies qualitatively and quantitatively. Merged label accuracy is shown quantitatively in table 1. Computational complexity is tabulated in Table 2. The same generated labels for the trajectory is applicable moreover across date, time etc.

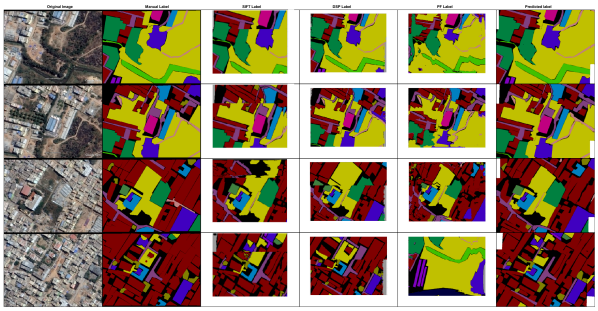

Fig. 7: Predicted merged labels. Column 1 and 2 original images with corresponding Manual labels. Column 3-5 and 6 have predicted labels from contemporary and proposed method

Methods	***mIoU***	***mAccuracy***
Siftflow [1]	84.1%	87.7%
DSPflow [2]	78.2%	82.5%
ProposalFlow [3]	81.3%	85.4%
Proposed	90.4%	93.5%

Table 1: Mean IoU and Accuracy

Methods	***Siftflow***	***DSPflow***	***PF***	***Proposed***
Time (in Sec)	6.8	9.3	48.7	1.6

Table 2: Typical Computational Complexity

To check the generalizability of the proposed approach, we extracted frames from the INRIA dataset [7] with corresponding labels at regular intervals. Additionally, compared with recent deep methods (SegNet and DeeplabV3+) for Inria and our dataset. The results are shown in table 3. We reduced the labels of our data set to building/non-building as available in Inria. For deep methods only inference is used. Performance of proposed method is of different order since it is a semi-automatic labeling approach in contrast with fully automatic deep methods.

Dataset	***SegNet [12]***	***DeeplabV3+ [6]***	***Proposed***
INRIA [7]	22%	19%	96%
Our dataset	24%	21.5%	92%

Table 3: Mean Acc. over standard data set and deep methods

The primary reason for the success of the proposed method is the fact that aerial videos taken from satellites or unmanned aerial vehicles flying at high altitudes primarily contain only ego motion. This global motion could have been good clue provided availability of synchronised data with accurate flight sensor (rarely available).

5 Conclusions and Future work

While dense flow is commonly employed for label transfer, it is expensive both in terms of memory and computation. We presented a light-weight label transfer method that exploited the sparse key point matching approach. In addition to IoU and accuracy parameters, we showed that the quality of transferred labels is comparable to the manual/ground truth labels. Recent deep and conventional methods explored on variety of our and INRIA data sets. We demonstrated improvement qualitatively and quantitatively with respect to state of art available in literature. Our label transfer approach paves the way for improved analysis of aerial imagery.

6 References

[1] Ce Liu, Jenny Yuen, and Antonio Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 978–994, 2010.

[2] Jaechul Kim, Ce Liu, Fei Sha, and Kristen Grauman, "Deformable spatial pyramid matching for fast dense correspondences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2307–2314.

[3] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce, "Proposal flow: Semantic correspondences from object proposals," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 7, pp. 1711–1725, 2018.

[4] Wei Yu, Xiaoshuai Sun, Kuiyuan Yang, Yong Rui, and Hongxun Yao, "Hierarchical semantic image matching using cnn feature pyramid," *Computer Vision and Image Understanding*, vol. 169, pp. 40–51, 2018.

[5] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8856–8865.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[7] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017, pp. 3226–3229.

[8] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[9] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European conference on computer vision*. Springer, 2004, pp. 25–36.

[10] David G Lowe et al., "Object recognition from local scale-invariant features.," in *iccv*, 1999, vol. 99, pp. 1150–1157.

[11] Fitsum Reda, Robert Pottorff, Jon Barker, and Bryan Catanzaro, "flownet2-pytorch: Pytorch implementation of flownet 2.0: Evolution of optical flow estimation with deep networks," 2017.

[12] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[14] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[15] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, vol. 202, pp. 18–27, 2017.

[16] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[17] Martin A Fischler and Robert C Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.