Large-Scale Mapping of Human Activity using Geo-Tagged Videos

Yi Zhu
University of California, Merced
yzhu25@ucmerced.edu

Sen Liu
University of Southern California
senliu@usc.edu

Shawn Newsam
University of California, Merced
snewsam@ucmerced.edu

ABSTRACT

This paper is the first work to perform spatio-temporal mapping of human activity using the visual content of geo-tagged videos. We utilize a recent deep-learning based video analysis framework, termed hidden two-stream networks, to recognize a range of activities in YouTube videos. This framework is efficient and can run in real time or faster which is important for recognizing events as they occur in streaming video or for reducing latency in analyzing already captured video. This is, in turn, important for using video in smart-city applications. We perform a series of experiments to show our approach is able to map activities both spatially and temporally.

CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**; *Video search*; • **Human-centered computing** → **Geographic visualization**; • **Computing methodologies** → **Activity recognition and understanding**; **Neural networks**;

KEYWORDS

Activity recognition, deep learning, convolutional neural networks, spatio-temporal analysis, geographic visualization

ACM Reference format:
Yi Zhu, Sen Liu, and Shawn Newsam. 2017. Large-Scale Mapping of Human Activity using Geo-Tagged Videos. In *Proceedings of SIGSPATIAL'17, Los Angeles Area, CA, USA, November 7–10, 2017,* 4 pages.
https://doi.org/10.1145/3139958.3140055

1 INTRODUCTION

Mapping human activity on a large scale in real time or near real time is a fundamental yet challenging task in the geographic and social sciences. It is an essential component for making cities smart, particularly with regard to resource allocation, disease control, social interaction, traffic management, etc. There has been work on using Twitter to geo-visualize human activity on maps [5]. There has also been work on using geo-tagged images to analyze human activity [2]. However, ours is the first work to exploit the rich temporal dimension of videos for activity mapping.

We therefore propose using geo-tagged videos to map human activity. We consider both the appearance and temporal (dynamic) aspects of the videos. This allows more effective activity detection than using tags/titles or the visual content of images.

Performing activity recognition in video is a challenging problem. Video data is large which makes real-time or near real-time analysis difficult. And, video data is very complex. Fortunately, the field of computer vision has made great progress recently in high-level video understanding thanks to deep learning. Large-scale labeled video datasets have been created, allowing deep convolutional neural networks (CNN) to be trained and achieve impressive performance on activity recognition. We take advantage of this recent progress to perform, for the first time, spatio-temporal mapping of human activity using geo-referenced videos.

This paper bridges activity recognition in video with geographic knowledge discovery. The salient aspects of the work include:

- Our work is the first to perform spatio-temporal mapping of human activity using the visual content of geo-tagged videos.
- We utilize an efficient video analysis framework termed hidden two-stream networks. The framework performs activity recognition at 130fps which allows it to run in real time.
- The video analysis framework is effective, achieving 90 percent accuracy on a 10 class activity classification problem.
- Our framework is flexible. It could easily be adapted to use geo-referenced videos to map a range of activities that are important for smart cities such as monitoring public safety, monitoring traffic, monitoring public health, etc.

2 RELATED WORK

Large-Scale Geo-Tagged Multimedia The exponential growth of publicly available geo-referenced multimedia has created a range of interesting opportunities to learn about our world. At the intersection of geographic information science and computer vision, large collections of geotagged photos/videos have been used to map world phenomenon [1], classify land use [17], model landmarks [8], conduct urban planning, and detect sentiment hotspots [19].

Our work is novel in that it uses a large collection of geo-tagged videos to map human activity as conveyed through the videos that ordinary people take. We specifically focus on spatial and spatio-temporal activity analysis in an urban area.

Visual Geo-localization Geo-localization is the problem of determining where something is. There exists an extensive body of literature on the large-scale visual geo-localization of images. Video geo-localization by comparison is relatively less studied. Note that our goal is not to perform geo-localization. Our videos are already geo-tagged. Our goal is to perform geographic knowledge discovery by analyzing the geo-tagged videos.

Video Activity Recognition The field of human action recognition in video has evolved significantly over the past few years.

SIGSPATIAL'17, November 7–10, 2017, Los Angeles Area, CA, USA

ACM ISBN 978-1-4503-5490-5/17/11.
https://doi.org/10.1145/3139958.3140055

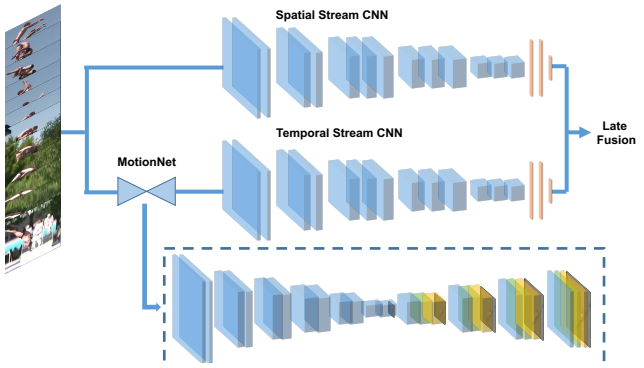

Figure 1: Illustration of the hidden two-stream networks that performs activity recognition using the visual content of a video. Both streams are end-to-end trainable.

Traditional handcrafted features such as Improved Dense Trajectories (IDT) [6, 11] dominated the field of video analysis for many years. Subsequent two-stream CNNs [7, 18] outperformed IDT by pre-computing optical flow and training a separate CNN to encode the motion information. However, pre-computing optical flow is computational and storage intensive and prevents traditional two-stream networks from running in real time. In this work, we utilize the recent hidden two-stream networks [16] for activity recognition. Our framework is extremely efficient yet maintains competitive accuracy with slower approaches which cannot operate in real time. We compare it for activity recognition with another state-of-the-art real-time activity model named C3D [10]. The results show the superiority of our method.

3 METHODOLOGY

The overarching goal of our work is to show that geo-referenced videos, such as at YouTube, can be used to spatio-temporally map human activity on a large scale. We select 8 popular sports, **baseball, basketball, football, golf, racquetball, soccer, swimming and tennis**, as common human activities to map. We also include the class **parade** to demonstrate how our approach can trace an event and the class **street fight** to show direct application to public safety. We thus consider 10 human activities in total but this could easily be extended to others. The fundamental technical problem we now face is human activity recognition in video. The next few sections describe our solution to this problem.

3.1 MotionNet

In order to achieve real time activity recognition, we use MotionNet [14, 16] instead of slower, handcrafted methods to compute optical flow. The key to using a CNN is to pose optical flow computation as a learning problem. MotionNet treats motion estimation as an image reconstruction problem [15, 20] where we seek to learn the optimal optical flow that allows the current video frame to be constructed from the previous one. Formally, given a pair of adjacent video frames I_1 and I_2 as input, MotionNet generates a motion field V. V and I_2 are then used to produce the estimate I_1' using inverse warping, i.e., $I_1' = \mathcal{T}[I_2, V]$, where $\mathcal{T}$ is the inverse warping function. The goal is to minimize the photometric (pixelwise) error between I_1 and I_1'. Training MotionNet to learn optimal optical flow involves minimizing the following three objective functions:

- A standard pixelwise reconstruction error function

$$L_{\text{pixel}} = \frac{1}{N}\sum_{i,j}^{N} \rho(I_1(i,j) - I_2(i + V_{i,j}^x, j + V_{i,j}^y)) \tag{1}$$

  where i and j are the frame numbers and V^x and V^y are the estimated optical flows in the horizontal and vertical directions. The inverse warping is performed using a spatial transformer module. We use a robust convex error function, the generalized Charbonnier penalty $\rho(x) = (x^2 + \epsilon^2)^\alpha$, to reduce the influence of outliers.
- A smoothness loss to address the ambiguity of estimating motion in non-textured regions (the aperture problem)

$$L_{\text{smooth}} = \rho(\nabla V_x^x) + \rho(\nabla V_y^x) + \rho(\nabla V_x^y) + \rho(\nabla V_y^y) \tag{2}$$

  where ∇V_x^x and ∇V_y^x are the gradients of the estimated flow field V^x in the horizontal and vertical directions. Similarly, ∇V_x^y and ∇V_y^y are the gradients of V^y. A generalized Charbonnier penalty $\rho(x)$ is also used.
- A structural similarity (SSIM) loss [13] is calculated as

$$L_{\text{ssim}} = \frac{1}{N}\sum(1 - \text{SSIM}(I_1, I_1')) \tag{3}$$

  where SSIM$(\cdot)$ is a standard structural similarity function. This forces MotionNet to produce flow fields with clear motion boundaries.

The overall loss is a weighted sum of the pixelwise reconstruction loss, the pixelwise smoothness loss and the region-based SSIM loss

$$L = \lambda_1 \cdot L_{\text{pixel}} + \lambda_2 \cdot L_{\text{smooth}} + \lambda_3 \cdot L_{\text{ssim}} \tag{4}$$

where λ_1, λ_2 and λ_3 weight the relative importance of the different metrics during training. λ_1 and λ_3 are set to 1.

3.2 Stacked Temporal Stream

Since MotionNet and the temporal stream are both CNNs, they can be stacked on top of each other and trained in an end-to-end manner. MotionNet takes consecutive video frames as input and outputs the estimated optical flow. The temporal stream CNN then uses this flow to predict activity labels. The stacked temporal stream CNN is later combined with a standard spatial stream CNN as shown in Figure 1. Following previous literature, the two streams are combined through weighted average late fusion using a spatial to temporal ratio of 1:1.5 as in [12].

4 EXPERIMENTS

4.1 Dataset

The dataset we use to train and validate our activity recognition model contains 10 activity classes. We only need the activity labels of these videos–we do not need geo-tags. We first leverage existing datasets including Sports-1M [4], UCF101 [9] and FCVID [3] to create an initial dataset. This initial dataset is too small and unbalanced though for fine-tuning deep CNNs and so we also download YouTube videos using the activity labels as keywords. Our final dataset contains 10,000 videos in total, 1,000 for each activity class. This size is of similar order to the UCF101 and ActivityNet 1.3

(a) Baseball (b) Basketball (c) Football

(d) Golf (e) Soccer (f) Tennis

Figure 2: Spatial mapping of popular sport in the city of San Francisco for 2016. (a) Baseball; (b) Basketball; (c) Football; (d) Golf; (e) Soccer; and (f) Tennis. Four detections are shown for each sport. This figure is best viewed in color.

datasets which have been shown to be large enough to fine-tune deep networks. We divide this dataset into training and validation components using a split ratio of 0.8:0.2.

To perform our spatio-temporal mapping, we download all geo-tagged YouTube videos using the same keywords within the city of San Francisco for the year 2016. This results in 265, 477 geo-tagged videos. Note that these videos are disjoint from the ones used to train and validate the activity recognition model above.

4.2 Activity Recognition Evaluation

We compare the accuracy and efficiency of our approach with the popular C3D network. The hidden two-stream networks achieves over 90.94% accuracy on the 10 class validation dataset at a speed of 130.56fps. It is about 6% more accurate than C3D (84.57% at a speed of 390.70fps). C3D is seen to be more efficient but both can run much faster than real time (30fps). The remainder of the experiments are performed with the hidden two-stream networks.

4.3 Spatial Sports Mapping

We now apply our framework to the geo-tagged YouTube videos from San Francisco for 2016. During inference, we sample frames every one second to reduce computational cost. Figure 2 shows the locations of videos classified as the six most popular sports. Also shown are four detections for each sport. We show a sample frame from the video that resulted in the detection as well as a satellite image of the location of the video. These results demonstrate that our approach is able to correctly classify the YouTube videos, and can use this classification to map where the activities take place.

Observation 1: Our approach is able to locate sports fields and complexes using the visual content of the geo-tagged videos. Figure 2(a) contains a concentration of points in the area of AT&T park, the home of the SF Giants baseball team. We also locate the San Francisco State University basketball court, George Washington High School football field, TPC Harding Park golf course, Crocker Amazon soccer fields, John McLaren Park tennis courts, etc.

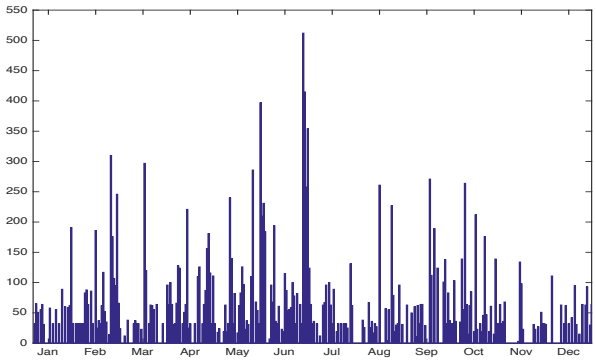

Figure 3: Temporal analysis of user uploaded parade videos in the city of San Francisco in year 2016. The y axis indicates the number of geo-tagged parade videos for each day. The peaks correspond to the major parades.

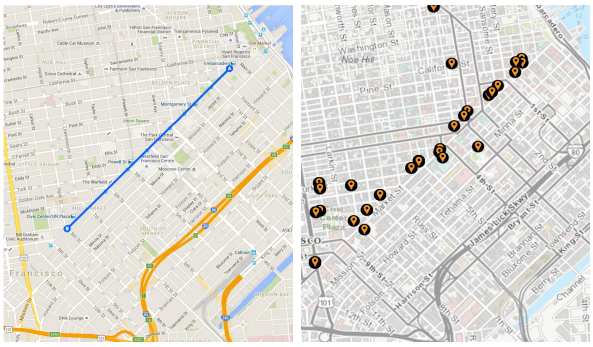

Figure 4: Spatial analysis of the 46th San Francisco Pride parade in 2016. L: official parade route. R: map of classified videos. Note the correlation.

Observation 2: The video frames and satellite images are in agreement with the predicted sports and their locations. There is, however, one interesting exception (the top right example in Figure 2(d)) of a golf video located in downtown. Upon further investigation, we found this makes sense since there is an indoor driving range inside the building named Eagle Club indoor golf. The classified video is an advertisement. This example demonstrates a distinct advantage that ground-level images have over satellite or aerial images–they can be used to perform geographic discovery indoors.

Observation 3: Our approach is able to use context to detect where a sport is played even if it is not occurring at the time the video was captured or the activity is difficult to discern. For example, in the top left example in Figure 2(c), the video snippet is an oblique view of just the football field. And, in the bottom right example in Figure 2(e), the players are very far from the camera. The ability of our approach to do this can be attributed to the spatial stream's capacity to learn the static appearance of where sports are played.

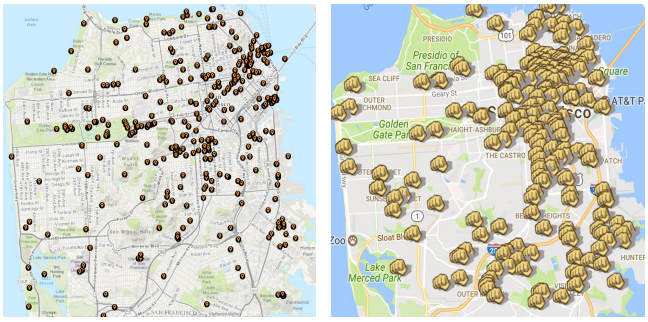

Figure 5: Violence detection. L: our predicted street fight mapping. R: official police record of Assault mapping.

4.4 Spatio-Temporal Parade Mapping

The goal here is to locate specific events, such as a parade, both spatially and temporally. We first detect all parade videos and temporally group them by date. We then map the videos in a group to identify the parade route.

Temporal Analysis: We detect a total of 15,645 parade videos in San Francisco in 2016. The daily distribution is shown in Figure 3. The peaks correlate with known parades including the Chinese New Year parade (February 20), the St. Patrick's Day parade (March 12), the Carnaval Grand parade (May 28), the Pride parade (June 25) and the Italian Heritage parade (October 9).

Closer analysis shows that the videos of a parade tend to be uploaded after the event, sometimes days later. This is different from texts or images which tend to be shared during the event. This is likely because video requires better network connectivity. Also, users often first edit their videos before uploading them.

Spatial Analysis: We now map the videos of the most popular parade in San Francisco in 2016 (based on our detections), the 46th Pride parade. As shown in Figure 4, our mapping results (right) are strongly correlated with the official parade route (left), from Market/Beale to Market/8th Street in downtown San Francisco.

4.5 Crime Detection

Detecting criminal activities is important for public safety. We here demonstrate how our framework can be used to map violence using YouTube videos. This shows how our framework can generalize to a range of applications related to smart cities given suitable training data. We apply our framework to the San Francisco YouTube videos and detect 7,784 instances of street fight. The locations of the videos are shown in Figure 5 left. We notice concentrations of violence in downtown San Francisco, the Mission District, Hunters Point, etc. These are known to be high-crime areas. For comparison, we show the locations of Assault from a San Francisco crime map in Figure 5 right derived from official police records. Our predicted locations are shown to be correlated with the official records.

We would like to point out how our framework is different and complementary to using traditional surveillance cameras to monitor crime. We use geo-tagged videos from YouTube. The challenge is that these videos are not taken from the same viewpoint, with the same camera, with controlled lighting conditions, etc. This makes our problem much more difficult. However, we are able to leverage the scale and embedded perspective of the crowd to detect incidents that might not be captured using surveillance cameras.

5 CONCLUSION

We performed the first investigation into using the visual content of geo-tagged videos to map human activity. We utilized the recent hidden two-stream networks to detect 10 different activities in a large collection of YouTube videos of San Francisco. Our approach can run in real time which is important for real world applications. In the future, we plan to investigate whether our framework can be adapted to detect a range of suspicious activities in surveillance video such as theft, vandalism, etc. Additional directions include scaling the mapping to country or continental regions as well as to more activity classes.

6 ACKNOWLEDGMENTS

We gratefully acknowledge the support of NVIDIA Corporation through the donation of the Titan X GPU used in this work. This work was funded in part by a National Science Foundation CAREER grant, #IIS-1150115, and a seed grant from the Center for Information Technology in the Interest of Society (CITRIS).

REFERENCES

[1] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. 2009. Mapping the World's Photos. In *WWW*.
[2] Eva Hauthal and Dirk Burghardt. 2016. Using VGI for Analyzing Activities and Emotions of Locals and Tourists. In *AGILE*.
[3] Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang. 2015. Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks. *arXiv preprint arXiv:1502.07209* (2015).
[4] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale Video Classification with Convolutional Neural Networks. In *CVPR*.
[5] Felix Kling and Alexei Pozdnoukhov. 2012. When a City Tells a Story: Urban Topic Analysis. In *ACM SIGSPATIAL*.
[6] Zhenzhong Lan, Yi Zhu, and Alexander G. Hauptmann. 2017. Deep Local Video Feature for Action Recognition. *arXiv preprint arXiv:1701.07368* (2017).
[7] K. Simonyan and A. Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. *NIPS* (2014).
[8] Noah Snavely, Steven M. Seitz, and Richard Szeliski. 2008. Modeling the World from Internet Photo Collections. *IJCV* (2008).
[9] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild. In *CRCV-TR-12-01*.
[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV*.
[11] H. Wang and C. Schmid. 2013. Action Recognition with Improved Trajectories. In *ICCV*.
[12] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. 2015. Towards Good Practices for Very Deep Two-Stream ConvNets. *arXiv preprint arXiv:1507.02159* (2015).
[13] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *TIP* (2004).
[14] Jason J. Yu, Adam W. Harley, and Konstantinos G. Derpanis. 2016. Back to Basics: Unsupervised Learning of Optical Flow via Brightness Constancy and Motion Smoothness. *arXiv preprint arXiv:1608.05842* (2016).
[15] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander G. Hauptmann. 2017. Guided Optical Flow Learning. *arXiv preprint arXiv:1702.02295* (2017).
[16] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander G. Hauptmann. 2017. Hidden Two-Stream Convolutional Networks for Action Recognition. *arXiv preprint arXiv:1704.00389* (2017).
[17] Yi Zhu and S. Newsam. 2015. Land Use Classification Using Convolutional Neural Networks Applied to Ground-Level Images. In *ACM SIGSPATIAL*.
[18] Yi Zhu and Shawn Newsam. 2016. Depth2Action: Exploring Embedded Depth for Large-Scale Action Recognition. In *ECCV Workshop*.
[19] Yi Zhu and S. Newsam. 2016. Spatio-Temporal Sentiment Hotspot Detection using Geotagged Photos. In *ACM SIGSPATIAL*.
[20] Yi Zhu and Shawn Newsam. 2017. DenseNet for Dense Flow. In *ICIP*.