ROBUST ACADEMIC POSTER RECOGNITION

Thanh Vu and Amir Sadovnik

Lafayette College
Department of Computer Science
Easton, PA, USA

ABSTRACT

Image recognition has been one of the most well researched tasks in the computer vision field. Many works have been published on recognizing pictures of buildings, works of art and other objects. Traditionally this is done by comparing features extracted from the query image to features extracted from an image database. However, although these produce state of the art results for many types of objects this type of approach does not generalize to all domains. In this paper we address one such domain: academic poster recognition. First we show that because of the unique structure of academic posters and the environment in which they are presented, the traditional approach of feature matching fails. Then we present a new approach for academic poster recognition based on object detection using convolutional networks and show how it outperforms the traditional approaches.

Index Terms— Poster Recognition, Object Recogntion

1. INTRODUCTION

Recent years have seen the integration of computer vision and machine learning algorithms in many applications. Automatic tagging of people in images, recognition of landmarks, and identifying different products using a mobile camera have become common applications. The advancements made in these fields both in accuracy and in speed have finally moved these algorithms from the research lab into user devices.

In order to move to a more applied domain there is usually a necessity to tailor the solution to that specific problem. For example, features that may do very well for landmark recognition might not be optimal for OCR. In addition, each one of these applications come with their own time constrains. While some are able to work offline and take a considerable amount of time, others might be useless unless they can run with very short delays. Therefore, different approaches have been proposed to deal with these constrains while optimizing performance.

However, there are still applied domains which will be of great use, but have not been well studied. In this paper we address one of them: academic poster recognition in conferences. The application for this domain is obvious. A robust and fast poster recognition system, will allow the development of applications from which a user can gain more information about a specific work simply by taking a picture of it. For example, once a photo of a poster is taken and recognized it can be linked to the original paper, a video describing the work, or other additional information.

Although it does have some similarities to other domains, academic poster recognition presents some challenges and some opportunities which are unique. For example, since many times a majority of the poster consists of text, it is not feasible to use a bag-of-words approach. Features extracted from the text will most likely match others from other posters. In addition, posters will usually be occluded by other conference attendees and only part of the poster might be visible. Finally, since the posters are custom made for each conference, we do not have many photos of the poster to train on. On the other hand posters are flat, their titles are usually at least partially visible even with occlusions, and we can easily obtain at least one digital image of the poster since they are usually designed on a computer. Examples of images taken at a conference are show in Fig. 1.

Fig. 1. In this work we attempt to develop a robust and fast algorithm to recognize academic posters in conferences by detecting and recognizing their titles.

In this work we attempt to develop a robust academic poster recognition algorithm which uses the simplifications this domain provides in order to address the difficulties. More specifically, in order for our poster recognition algorithm to be practical we make the following assumptions:

- The algorithm must be relatively fast and it's running time should not grow significantly with the number of posters it can recognize.
- In the photos taken, the title of the photo should be at least partially visible. From our observations the posters themselves tend to be occluded by viewers in most photos. However, since the titles are usually the highest part of the poster they tend to be mostly visible.
- In order for the algorithm to work in real conferences it should be able to train on one image of the poster (the original PDF) as opposed to many images of the same poster. We do not consider training time to be a critical issue.

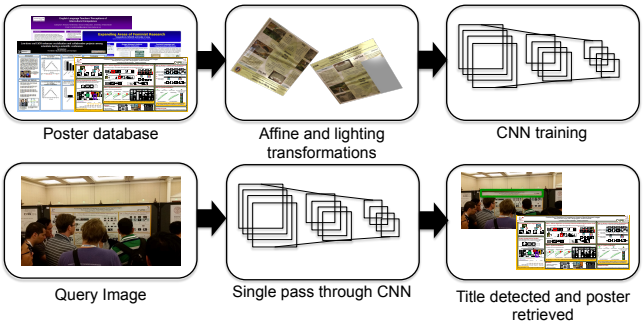

Fig. 2. The general flow of out method. On top we show our training. For each poster in our database we generate a set of images using different transformations (Sec. 3). We then use these newly generated images to fine tune the weights of an object detection CNN. On the bottom we show our testing. Given a query image we simply pass it through our trained CNN. We then look at the posters detected in the image, and use the one with the highest score as our poster label.

2. PREVIOUS WORK

Although as far we know there has been no previous work specific for academic poster recognition, there is obviously a vast amount of work on object recognition in images. One of the most common ways of achieving this goal in the past was to first extract specific invariant features. Different features have been proposed (such as SIFT [1], ORB [2], SURF [3], etc.) where the differences between these features are mostly in their invariance, their uniqueness, and the speed in which they are extracted and matched. Once an image can be described in terms of its features, the algorithm attempts to match features from one image to another. This can be done using a simple bag of words approach of matching histograms [4], or more advanced matching which includes spatial information such as spatial pyramids [5] or geometry preserving visual phrases [6].

This specific pipeline has been used for recognizing images in many different domains. Some of the more common examples are landmark and building recognition [7, 8] and painting recognition [9]. However, since these features do not work for all possible domains other descriptors have been proposed which address specific types of images. For example, Friedman et al. [10] propose recognizing icons using a custom engineered descriptor which works better than more generic features. As we show in Sec. 4, since the more generic features do not work well for academic posters we wish to explore a new representation which would achieve higher accuracy. However, instead of engineering our own feature representation we attempt to use convolutional neural networks to find an optimal representation.

3. METHOD

Our method is visualized in Fig. 2. First, we train a convolutional neural network based on the YOLOv2 network [11, 17]. We treat each poster title as an object class and train the network to detect the titles in an image. However, since training a network of this size requires many images per class, we first need to generate synthetic data. We do this in a similar manner to [12]. Finally, after we have trained the network we can use it on a new image to detect which poster is in the image by simply passing the image once through our neural network.

3.1. The Yolo Network

As mentioned in Sec. 1, our goal is to find a poster recognition system that is both fast and robust. As has been shown in previous years, convolutional neural networks have managed to achieve state of the art results on many tasks, and therefore seem to be a natural choice. However, traditionally these networks have been very slow especially for the object detection task. For example, approaches like R-CNN [13] require multiple steps for recognition. First region proposals must be made to generate potential bounding boxes. Then each of these regions must be passed through a large CNN to extract features. Finally these features are scored by a class specific SVM (one per class). Although there have been faster implementations of the R-CNN [14, 15] they still are relatively slow because of the pipeline needed.

However, the network proposed in [11] is able to perform robust detection very fast because it unifies all stages into one CNN. The general idea is to divide the image into an $n \times n$ grid and output a probability that an object's bounding box is centered at that location. Thus detection of multiple objects can occur with just one pass through the neural network. We therefore start with the trained YOLOv2 network, and try to fine tune the weights for recognizing posters.

3.2. Data Synthesis

We assume that we only have one image of the poster originally (for example the original PDF). Since we are treating each poster as an object class we need many more images to fine tune the weights in the YOLOv2 network. Luckily, since we are dealing with posters we can make the assumption that they are planar and therefore will only undergo certain transformation. We can automatically generate new images of the posters after simulating these transformations and then train on our newly generated image set. Our goal in applying these transformations is to try and capture the distortions that a poster might undergo when its photo is taken at a real live conference.

For this task we use the following transformations:

3.2.1. Occlusions

We observed that in certain contexts such as a research conference, images of posters have considerable amount of occlusion due to the audience. Hence, as mentioned in Sec. 1 we focus our learning on the titles, which are much more likely to be captured with a minimal amount of blockage, and use them as our object to be detected. We let the titles width to be equal to that of the poster itself, and estimated the titles height to be one fifth of posters height.

With that assumption, this step simulates the poster's audience (occlusions) as gray rectangles. The number of generated rectangles is random and within the range of 0 to 10. The location of each occlusion is also randomly assigned within the dimensions of the images. Note that to ensure training efficiency, we generated occlusions to block parts or all of the posters body but not the title.

3.2.2. Lighting Transformation

Since posters tend to be glossy and reflect light, we noticed that many of our poster images would have bright spots on them (reflecting different light sources in the room). We simulate this by generating light blobs. The number of blobs is randomized from 0 to 4. The light blobs are created using a Gaussian distribution with arbitrary mean (randomly in the poster) and a standard deviation =

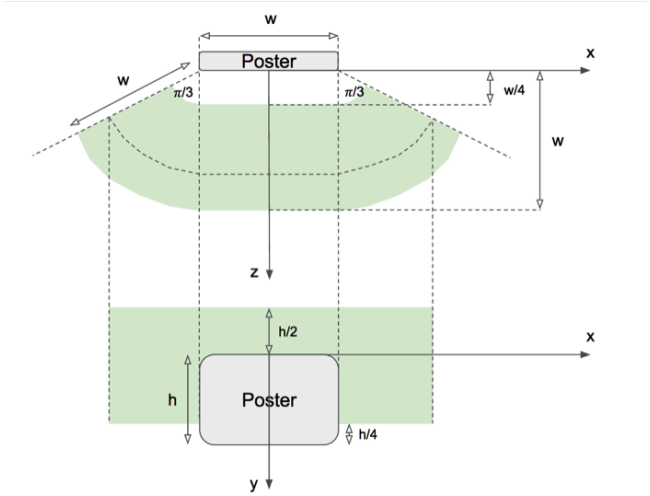

Fig. 3. 3D point selection for the perspective transformation. We choose a random point in the green area which has been deemed to be a reasonable location to take a picture from.

$R/3$, where R is randomized between $h/2$ and $h * 2$. These blobs are then placed randomly within the title's area.

3.2.3. Blurring

We simulate the fact that images are occasionally taken out of focus. We do this by applying a standard average blurring method on the training images. More specifically blurring is done by performing convolution on the image with a kernel K, where:

$$K_{n,n} = \frac{1}{n^2} \times \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix} \quad (1)$$

We choose a random n between 4 and 9.

3.2.4. Scale, Translation and Rotation

In this step we perform different 2D transformations. First, we scale the poster by varying its size, relative to the training image itself. In other words, the ratio of a training image and the poster is randomized between 2:3 and 1:1. After being scaled the poster is translated within the training image. While translating, we also allow for the possibility that a fraction of the poster might get cropped out. This helps simulate the fact that a test image taken by a user might accidentally not capture the entire poster. Finally, the poster is rotated by a random angle between $-\pi/4$ and $\pi/4$. We also experiment with completely randomizing the rotation between 0 and 2π.

3.2.5. Perspective Transformation

In addition to the 2d transformations discussed, we need to perform 3d transformations. This is done by performing a perspective transformation. First, we randomly choose a viewpoint, which simulates the location of the camera at the time of taking pictures. The point is generated randomly within the 3D area illustrated in Fig. 3 (green area). The maximum distance to the poster is w (where w is the width of the original poster), and the view angle is between $-\pi/3$ to $\pi/3$. The height of the point is the upper $3/4$ of the poster plus an additional $h/2$. These were estimated to be a reasonable viewpoint range. We then pass the generated viewpoint to a 4×4 projection matrix to create a 3D perspective transformation of the poster. We use the Homogeneous Transformation Matrices and Quaternions Library [16] to do so. Finally, we convert the 3D transformation to 2D to get the final training image.

Fig. 4. Examples of our results. The top two images show successful detections and classifications, while the bottom two show failed detections. This obviously lead to failed classification as well

3.3. Fine-Tuning the Yolo Network

Once we have generated our training set we can use it to fine tune the weights in the YOLOv2 network. We do this by first changing the amount of nodes in the final convolutional layer depending on the amount of posters we are training on. More specifically, in accordance with [17] the final layer needs to have $f = (p + 5) \times b$ where f is the number of nodes, p is the number of posters (classes), 5 is the number of coordinates returned for each box (4 to describe a bounding box and one for the confidence level), and finally b is the number of boxes which in our case is 5. We then train the network with our synthesized dataset using stochastic gradient descent.

4. DATASET AND RESULTS

4.1. Academic Poster Dataset

We collect a dataset of 100 academic posters from different departments at Lafayette College. For each poser we first take a front facing photo which includes the entire poster (and only the poster) oriented correctly and then we take 5 additional photos of the poster for testing. The test images are varied in the way they are taken. This includes taking photos at different distances, angles, cropping and occlusions. All photos are taken using a mobile device .

We use our single training image (per poster) to generate the training set for the neural network by performing the image transformation described in Sec. 3. For each training image we create a set of 2000 images using random transformations. We then test on the five remaining images for each poster.

4.2. Classification Results

Once we have our YOLOv2 network trained for detecting the titles of the posters we can begin the test phase. Given a query image we

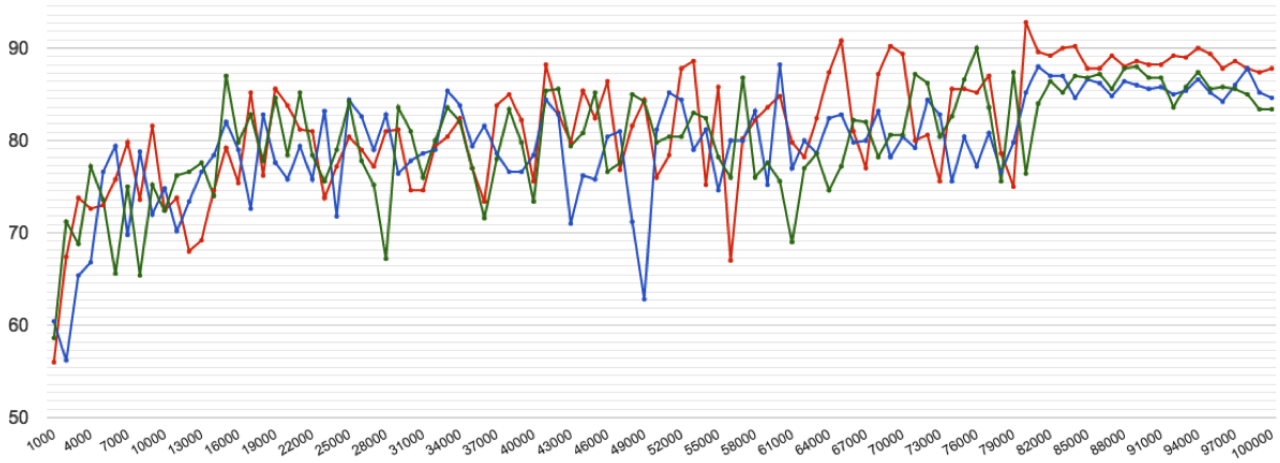

Fig. 5. The improvement in our network detection as a function of time. We show results with all the transformations (red), results without adding lightning transformations (blue) and results without adding blurring transformations (green). Using all transformations seems to yield the optimal results.

# Of Posters	SIFT	ORB	YOLO2
20	84.33	77.33	97.0
50	79.6	71.07	91.6
100	76.0	66.6	90.8

Table 1. Results using different number of classes (posters).

20 Posters			100 Posters		
Iteration	Time(hr)	Acc.	Iteration	Time(hr)	Acc.
0.9k	2	90.0%	19k	36	85.6%
10k	24	97.0%	65k	132	90.8%
34k	36	100.0%	80k	168	92.8%

Table 2. Training time for different amounts of posters (classes) and accuracies .

run it through our neural network. The output of the network will specify which poster titles have been detected. We then classify the image as being an instance of the poster which title was detected with the highest confidence level. If no title was detected, we simply do not classify this image as any poster. Since we only use the original poster for training (plus our transformations) we can use the 5 additional posters as query images and measure our accuracy as the percentage of posters which were classified correctly.

In order to present the effectiveness of our results we compare it to a simple baseline. More specifically we use the traditional feature matching approach. For each training poster we extract a list of feature vectors. Then given a test image we extract the same type of feature vectors and find the nearest neighbor for each one of them. Finally, for each training image we count how many features were the closest to the test image, and choose that as our match. We use two common features to compare to: SIFT [1] and ORB [2].

We present our results in Table 1. We conduct experiments with three different amounts of original posters (classes). As can be seen in the table, our method clearly outperforms both the baselines by a wide margin. In addition, our method is able to perform well even when faced with a large number of posters (92.8% accuracy with 100 posters). Examples are shown in Fig. 4.

The method is efficient during test time since it only needs to pass the image through the neural network once (0.13s per image). However, training is relatively slow. As shown in Table 2 the network requires around 65k iterations to reach reasonable results for the 100 posters scenario. These 65k iterations take 5 days using a single Nvidia Tesla K40. Although training is slow we do not consider this to be a bottleneck since it can be done ahead of time, and training time would also improve using faster GPU's and parallel training.

In addition we also wish to examine how some of the transformations we proposed effect the final results. We focused specifically on blurring (Sec. 3.2.3) and lighting (Sec. 3.2.2) since their effect is not as obvious as some of the other geometrical transformations. We compare the results without lighting and without blurring in Fig. 5. Although these transformations do not improve the results by much, our final results using both transformations achieves about 2% better classification then not using one of them.

5. CONCLUSION

In this paper we develop a fast and robust academic poster recognition algorithm. Many of the traditional matching techniques use features which are not optimal for these types of images. In addition, in real life settings many posters would be partially occluded in an image. Therefore we propose an algorithm based on object detection using deep convolutional neural networks which we predict will be better since they learn the features from the data. We take advantage of the fact that academic posters are planar and propose a set of transformations in order to generate a large training set for the convolutional neural network. We then show through experiments how our method outperforms more traditional baselines.

6. REFERENCES

[1] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, 2004.

[2] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*, 2011.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*, 2006.

[4] Christian Wallraven, Barbara Caputo, and Arnulf Graf, "Recognition with local features: the kernel recipe," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003.

[5] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006.

[6] Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen, "Image retrieval with geometry-preserving visual phrases," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011.

[7] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven, "Tour the world: building a web-scale landmark recognition engine," in *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on*, 2009.

[8] David M Chen, Georges Baatz, Kevin Köser, Sam S Tsai, Ramakrishna Vedantham, Timo Pylvänäinen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, et al., "City-scale landmark identification on mobile devices," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011.

[9] Niki Martinel, Christian Micheloni, and Gian Luca Foresti, "Robust painting recognition and registration for mobile augmented reality," *IEEE Signal Processing Letters*, 2013.

[10] Itamar Friedman and Lihi Zelnik-Manor, "Icon scanning: Towards next generation qr codes," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012.

[11] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[12] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman, "Synthetic data for text localisation in natural images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[14] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015.

[16] Christoph Gohlke, "Homogeneous transformation matrices and quaternions," http://www.lfd.uci.edu/ gohlke/, 2006-2015.

[17] Joseph Redmon and Ali Farhadi, "Yolo9000: Better, faster, stronger," 2016.