Jersey Number Recognition

Computer vision course project

Alaa Hesham

Zewailcity of Science and Technology
s-alaahesham@zewailcity.edu.eg

Omar Yasser

Zewailcity of Science and Technology
s-omaryasser@zewailcity.edu.eg

***Abstract*—recognizing jersey number of players in soccer matches is a very challenging task to achieve till now, in this paper we will apply various methods such as Maximally stable extremal regions (MSER), template match , Scale-invariant feature transform (SIFT), and you only look once (YOLO)network. Results of every techniques, and how to improve them will be clearly shown.**

Keywords— Jersey numbers; MSER; template match; SIFT; YOLO.

I. INTRODUCTION

To support automatic understanding for soccer match videos, recognizing jersey numbers is very important and required task to identify players since every player in his team has a distinct number from 0 to 99. Hence, if distinguishing between two teams has been achieved, then we managed to identify jersey number for every team player, we will manage to have data necessary to analyze soccer matches to a great extent [1] .At the first glance, the task seems really easy as number and text recognition are considered as essential and classical problems in computer vision field [2, 3] .However , Jersey number recognition is still challenging due to camera perspective , soccer video resolution ,motion blur ,and light illumination [1] .In this paper , we will demonstrate results obtained after trying many algorithms while approaching the problem at hand. These algorithms are maximally stable extremal regions (MSER), scale-invariant feature transform (SIFT), template matching, and optical character recognitions (OCR) such as Yolo object detection system.

II. METHODOLOGY

To distinguish between two teams, we used the following:

We apply a color mask that selects one team and rejects the other .For example, if we are willing to detect team whose t-shirts are in red then applying a mask that detect red color is a good idea, and also we use another feature to make sure that we detect humans which is human always have height bigger than width.

To recognize jersey numbers, we used the following:

A. MSER

MSER is a feature detector algorithm which extracts covariant regions from image "I" [4] called MSERS .Every MSER is a stable connected component of some level sets of the image "I". In other words, if an image has been fed to MSER algorithm, the result will be patches from it where every patch represents a coherent region. Every region is quite different from other regions. The reason that we have chosen it to be our first algorithm to try is that usually jersey number has high contrast with players' shirts in order to be easily recognizable so it will be one of regions that MSER will detect. If we feed results to OCR to detect patches that contain numbers, we would expect to detect jersey numbers in the image.

B.Template match

Template match is a technique that matches particular regions of a certain image to a template image or patch .In other words, it requires the existence of a source image and template image. It slides the template image over source image pixel by pixel. A metric is calculated to represent the accuracy or how good the matching happens. Then the results of matching is saved into matrix called *result matrix,* in which every entry represent the metric match at this location (x,y) [5]There are six methods

a. method=CV_TM_SQDIFF

$$R(x,y) = \sum_{x',y'} (T(x',y') - I(x+x',y+y'))^2$$

b. method=CV_TM_SQDIFF_NORMED

$$R(x,y) = \frac{\sum_{x',y'} (T(x',y') - I(x+x',y+y'))^2}{\sqrt{\sum_{x',y'} T(x',y')^2 \cdot \sum_{x',y'} I(x+x',y+y')^2}}$$

c. method=CV_TM_CCORR

$$R(x,y) = \sum_{x',y'} (T(x',y') \cdot I(x+x',y+y'))$$

d. method=CV_TM_CCORR_NORMED

$$R(x,y) = \frac{\sum_{x',y'} (T(x',y') \cdot I(x+x',y+y'))}{\sqrt{\sum_{x',y'} T(x',y')^2 \cdot \sum_{x',y'} I(x+x',y+y')^2}}$$

e. method=CV_TM_CCOEFF

$$R(x,y) = \sum_{x',y'} (T'(x',y') \cdot I(x+x',y+y'))$$

where

$$T'(x',y') = T(x',y') - 1/(w \cdot h) \cdot \sum_{x'',y''} T(x'',y'')$$
$$I'(x+x',y+y') = I(x+x',y+y') - 1/(w \cdot h) \cdot \sum_{x'',y''} I(x+x'',y+y'')$$

f. method=CV_TM_CCOEFF_NORMED

$$R(x,y) = \frac{\sum_{x',y'} (T'(x',y') \cdot I'(x+x',y+y'))}{\sqrt{\sum_{x',y'} T'(x',y')^2 \cdot \sum_{x',y'} I'(x+x',y+y')^2}}$$

Figure 1 Different metric options to measure distance

We have tried six different methods to come up with the most suitable one. The reason that we choose this algorithm to be second one to try is that if we have many patches for every player, we can detect these players if any of these patches have appeared.

C.SIFT

The scale-invariant feature transform (SIFT) is an algorithm used to detect and describe local features in digital images. It locates certain key points and then seek to describe them (so-called descriptors). The reason we choose it is that SIFT is invariant for scale, rotation, clutter, and occlusion as it depends on local features rather than global ones, so we expect that we will not have to provide SIFT algorithm with great number of patches as template match as it did not require that patch to be exactly the same to be able to match it due to its properties.

D. Yolo

Yolo is a deep learning model that works as an OCR, it is trained on a dataset that detects house numbers on the streets, we find it is a good candidate as it detects numbers against different backgrounds, we have experimented feeding the scence and let it recognize numbers itself.

III. RESULTS AND DISCUSSION

Regarding distinguishing between two teams, we have reached satisfying results:
It correctly classifies two teams, however in some body positions it could not classify them.

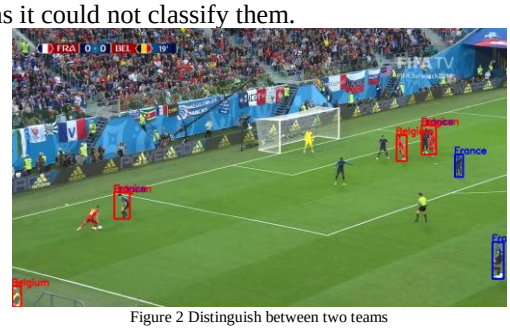

Figure 2 Distinguish between two teams

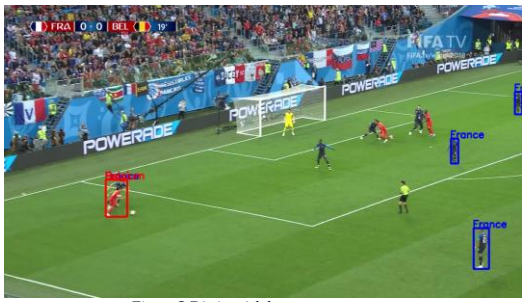

Figure 3 Distinguish between two teams.

Regarding recognizing jersey numbers, we used the following:

A.MSER

For MSER, when we use it to detect numbers on image of match scene, the results were like the following:

Figure 2 Source image

Figure 3 Resulted Patches

The main problem is that the patches include both numbers, and non-numbers .For numbers, we have multiple numbers here so we cannot tell if there are, for instance in that scene, players whose numbers are 1, 3, 13 or just one player whose number is 13. There are proposed solutions in [6, 7], which is to use CNN that has three classifiers. First one is to detect first digit, second one is to detect second digit, and third one is to detect length of the sequence to determine whether it is composed of one digit or two. In that example it will determine if we have three players whose numbers are 1, 3, 13 or just one player whose number is 13.The main problem is that this algorithm depends so much on predicting length, and predicting length mainly depends on camera perspective variations which is the case on soccer match videos, that is why, we have switched to another technique, template match.

B. Template match

This algorithm has a main advantage which is if it takes a given patch mentioned in the scene , it will detect it correctly . However , the main disadvantage is that it should be exactly the same patch otherwise it will not detect it so you need to provide it with great number of patches for every player .

Figure 4 Patch (Template)

Figure 5 Big image

That is why we have thought that it could be a good idea to use SIFT as it is scale, rotation, clutter, and occlusion invariant. So we could provide it with smaller number of patches to match them as patches could be of different sizes, rotated, occluded, and cluttered, and still be detected.

Template match has six metric methods in opencv , the one that has worked the best was *CV_TM_CCOEFF_NORMED*.

C.SIFT

SIFT's results were unlike expected ones. If the patch was exactly the same as its corresponding part, it will detect it correctly just like Template match.

Figure 6 SIFT results while using identical patch from scene

If the patch was slightly different, for instance belongs to the next frame of soccer match video, it will result in false positive results.

Figure 7 SIFT results while using slightly different patch

We have realized that the results are the way they are as SIFT is variant to deformation which happens in soccer matches is deformation not rotation as SIFT is invariant to rotation up to 30 degrees only .Also interest points in SIFT depends on the existence of corners and severe changes but numbers are quietly look like each other .That is why SIFT performs poorly to recognize numbers.

D. *Yolo*

Although Yolo is an OCR which is trained to detect numbers of houses , it performs poorly on detecting jersey numbers in soccer matches , the reason is that it is trained on dataset that detect house numbers in the streets , results could be way better if the training dataset was for jersey numbers .

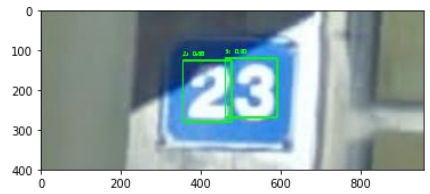

Figure 8 Yolo correctly detects houses' numbers

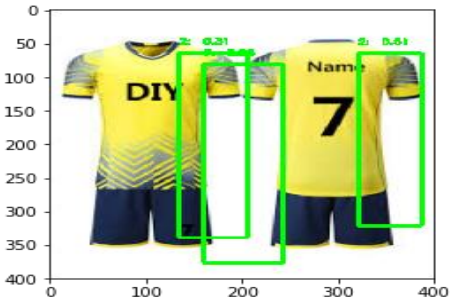

Figure 10 Yolo wrongly detects soccer numbers

IV. FUTURE ENHANCEMENT

A. *Multi-scale patches in template match*

Instead of having many patches, we can use less number of patches yet for every patch we can provide many scales in order to be able to detect it when the camera angle is far or near. In other words, make template match scale invariant.

B. *Dataset*

Training Yolo with dataset of soccer numbers is expected to result in great outcomes .A Dataset called SJN-210K is developed by [1] will supposedly be available for research soon. We expect that Yolo will be able to perform on jersey numbers as good as on houses' numbers.

C. *Image segmenation*

Since we managed to recognize where players are, we can just use regions where they exist to search for jersey numbers instead of the whole image.

V. CONCLUSIONS

This project has two main goals:

First one is to distinguish between two teams' players, this goal has been achieved to a great extent, there are certain body positions where it becomes difficult to distinguish between players yet generally results are satisfying.

Second one is to recognize jersey numbers, we have tried several techniques. Every technique has its own drawbacks, and advantages as explained in the results section .At this point, we see that template match is the one which gives the most accurate results but unfortunately it requires a lot of patches to work on them. We see that Yolo could result in promising outcomes if it has trained on SJN-210K dataset, and MSER could perform better if it was followed by motion tracking system that determines which number is really on the image before we feed resulting patches to OCR.

REFERENCES

[1] G. Li, S. Xu, X. Liu, L. Li and C. Wang, "Jersey Number Recognition With Semi-Supervised Spatial Transformer Network", Openaccess.thecvf.com,2019.[Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018_workshops/w34/html/Li_Jersey_Number_Recognition_CVPR_2018_paper.html. [Accessed: 10- May- 2019].

[2] C. Bartz, H. Yang, and C. Meinel. See: Towards semisupervised end-to-end scene text recognition. arXiv preprint arXiv:1712.05404, 2017.

[3] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven. Photoocr: Reading text in uncontrolled conditions. In Computer Vision (ICCV), 2013 IEEE International Conference

[4] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In BMVC, 2002 .

[5] "Template Matching — OpenCV 2.4.13.7 documentation", Docs.opencv.org, 2019. [Online]. Available: https://docs.opencv.org/2.4.13.7/doc/tutorials/imgproc/histograms/template_matching/template_matching.html. [Accessed: 10- May- 2019].

[6] S. Gerke, K. Muller, and R. Schafer. Soccer jersey number recognition using convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 17–24, 2015.

[7] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082, 2013. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.