The Hague, 18/04/2019

Applying local image feature descriptors to aid the detection of radicalization processes in Twitter

This paper was presented at the 2nd European Counter-Terrorism Centre (ECTC) Advisory Group conference, 17-18 April 2018, at Europol Headquarters, The Hague.

Authors: Daniel López-Sánchez, Juan M. Corchado

1 Introduction

With the emergence of the online-radicalization phenomenon, several researchers have turned their attention towards the problem of automatic detection of online radicalization processes and profiles. In addition, several studies have been conducted trying to provide insight into the habits and language patterns used by radical profiles to spread their radical ideology. For instance, in [4] the authors manually identified a set of radical YouTube profiles and used different social network and natural language processing techniques to analyze the messages they published in the social network. This study revealed significant gender differences in the language and habits of radical users.

Other authors have focused on the automatic detection of radical profiles. Most of these proposals focus solely on the textual content of individual publications, rather than analyzing the social interactions between confirmed radical users and users at risk of radicalization. For example, in [1] the authors adopted a machine-learning classification approach to detect ideologically extremist tweets based on linguistic and stylistic text features.

In recent years it has become apparent that, to fully characterize the behavior of radical social network users, it is necessary to consider not only the textual content of messages but also the patterns in the interactions between social users. It has been shown that users in social networks interact in a homophilic manner; that is, they tend to maintain relationships with people who are similar to themselves, as characterized by age, race, gender, religion and ideology. For instance, in [6] the authors analyzed different community detection techniques to cluster users according to their political preferences, showing that users in the social network Twitter tend to form very cohesive networks when talking about political issues.

In this context, our previous work [7] focused on the design of algorithms to measure the risk of radicalization of social users surrounding networks of confirmed radical users. If accurate enough, these algorithms might be applied by security forces to detect early radicalization processes, allowing the adoption of effective counter-measures in a timely manner. As described in the following section, our proposed algorithm was able to correctly identify users at risk of radicalization in different case studies. However, one limitation of the proposed framework was the high level of false positives that expert users had to filter out manually. In this document, we explore the use of image analysis algorithms in order to detect radical groups' iconography in social network profile images and publications. We believe this technology can be used in combination with existing

interaction and text-based radical user detection techniques in order to reduce false positive rates. In particular, the presence of iconography from radical groups in online publications can be used to confirm or at least support the predictions of existing radicalization detection methods, prioritizing users which share or exhibit iconography of radical groups in their social profiles.

The rest of this document is structured as follows. Section 2 briefly summarizes our previous work in the context of interaction-based radicalization detection, and explains how image processing techniques can improve the accuracy in this task. Section 3 presents some experimental results regarding the effectiveness of different image descriptors when retrieving images containing iconography of radical groups. Finally, section 4 presents the conclusions of this study and suggests some promising future lines of research.

2 Using image feature descriptors to assist the detection of radical profiles online

As mentioned before, our previous work in the field of radicalization detection [7] focused on the analysis of social network interactions between confirmed radical users and users at risk of radicalization. Initially, the monitored network consists of radical users identified either by expert knowledge or automated tools. More specifically, we have a set of radical users and a quantitative measure of their radicalization influence level:

$$U = \{(u_1, r_1), (u_2, r_2) \ \ldots, (u_n, r_n)\} \quad (1)$$

Where u_i is a radical user and $r_i \in [0,1]$ a measure of his/her radicalization influence. This influence can be assigned manually or estimated as a function of the number of followers, retweets and favourites. In particular, we proposed calculating the radicalization influence of a given user u_i as follows:

$$r_i = \min(1, \frac{RTcount + FavCount}{tweetsCount}) \quad (2)$$

Then, our approach analyzed the social interactions with surrounding users to measure their risk of radicalization. Once the monitoring process began, all the previously published tweets of monitored users were downloaded and analyzed. The goal here is to find interactions (e.g. mentions and retweets) between radical users and potentially vulnerable users of the social network. To this extent, the following information is considered:

- The list of all mentions published by the monitored users: $M = \{M_1, \ \ldots, \ M_m\}$
- All retweets done by the monitored users: $RT = \{RT_1, \ \ldots \ , RT_r\}$

Every user that has interacted with any of the monitored radical profiles is analysed and his/her risk of radicalization is estimated. The risk of a given user u being radicalized is computed as follows:

$$Risk(u) = \sum_{i=1}^{n} r_i \cdot follows(u, \boldsymbol{u_i}) + \sum_{i=1}^{n} r_i \cdot follows(\boldsymbol{u_i}, u) + \sum_{i=1}^{r} RT_i.to(u) + \sum_{i=1}^{m} M_i.to(u) \cdot |sentiment(M_i)| \quad (3)$$

where $|sentiment(M_i)| \in [0, 1]$ is the automatically estimated absolute value of sentiment of the mention's original text [10]. Figure 1 shows an example of how this formula is applied in practice.

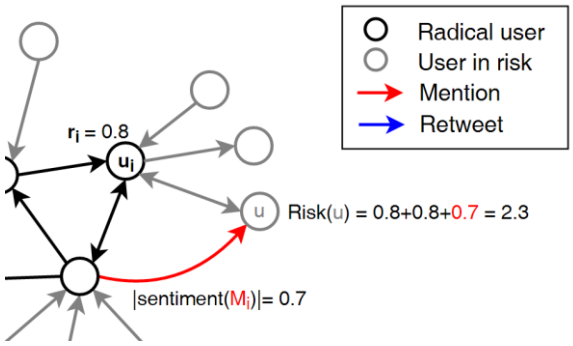

Fig. 1. Example of radicalization risk estimation

While this approach was able to detect users at risk of radicalization in several case studies, we noticed that a major drawback of our method was the potentially high level of false positives. This mainly happened for profiles which were frequently mentioned by the radical users but never responded to those mentions (e.g., mentions to politicians who were profusely criticized by the radical users under monitoring). To mitigate this issue, we propose applying image analysis techniques to prioritize in the list of users at risk of radicalization those profiles which share or exhibit iconography of radical groups. In particular, this approach requires the expert user managing our framework to provide a reference image containing the characteristic iconography of the corresponding radical group. Then, different image descriptors can be applied to determine the presence of this iconography in the images shared by monitored users in social networks (e.g., profile images or public posts). In particular, in our experiments we compared four different descriptors:

- Scale-Invariant Feature Transform (SIFT), introduced in [8], is arguably one of the most effective and widely used local image descriptors. The major drawback of this algorithm is its computational cost, being significantly slower than alternative methods.

- Root-SIFT, popularized by [2], is a simple extension of the SIFT descriptor which can potentially boost the results by simply L1-normalizing the SIFT descriptors and taking the square root of each element.
- Speeded Up Robust Features (SURF), proposed in [3], was conceived as a faster alternative to SIFT which sacrificed little or none of the accuracy of its predecessor. Nowadays, SURF is almost as popular as SIFT, and the results for both methods are comparable with the exception that SURF runs significantly faster.
- Oriented FAST and Rotated BRIEF (ORB), introduced more recently [9], is a computationally efficient alternative to SIFT and SURF. In addition, as opposed to SIFT and SURF, ORB is not patented and, thus, is free to use. The major difference with SIFT and SURF is that ORB uses binary descriptors to achieve its remarkable efficiency.

The mentioned algorithms are applied in the following manner to detect the presence of radical iconography in images:

1. First, each image is analyzed to detect the key points. These are points in the image with a characteristic visual appearance than can be useful to detect similar images. In this regard, each compared descriptor has its own key point detection strategy and more information can be found in the corresponding references.
2. Secondly, a feature descriptor is generated for each key point. The descriptors are vectors of fixed length, whose contents describe the appearance of the previously detected key points.
3. Then, to determine the presence of the selected radical iconography in a target image, the descriptors of both the reference and the target images are compared, finding the best matches in terms of Euclidean distances (or normalized Hamming distances for ORB) between the descriptors. Valid matches were determined using the ratio test proposed by D. Lowe in [8]with a 0.8 threshold.
4. The predicted probability for a target image of containing radical iconography is computed proportionally to the number of descriptor matches between the target and the reference images.

The following section contains experimental results comparing the accuracy of these descriptors in the task of radical iconography detection.

3 Heading Experimental protocol and results

To evaluate the accuracy of the different image descriptors considered in the previous section, we collected a dataset of real-word occurrences of radical iconography. In particular, we selected three radical groups with characteristic iconography. Namely Golden Dawn (GD), a Greek ultranationalist party; Islamic State (ISIS), a jihadist terrorist organization and Hogar Social Madrid (HSM), a neo-fascist group. First, we selected a representative image for the iconography of each group. These are the images that the human experts might provide to our tool as the reference. The selected reference images are shown in figure 2, alongside their detected key points.

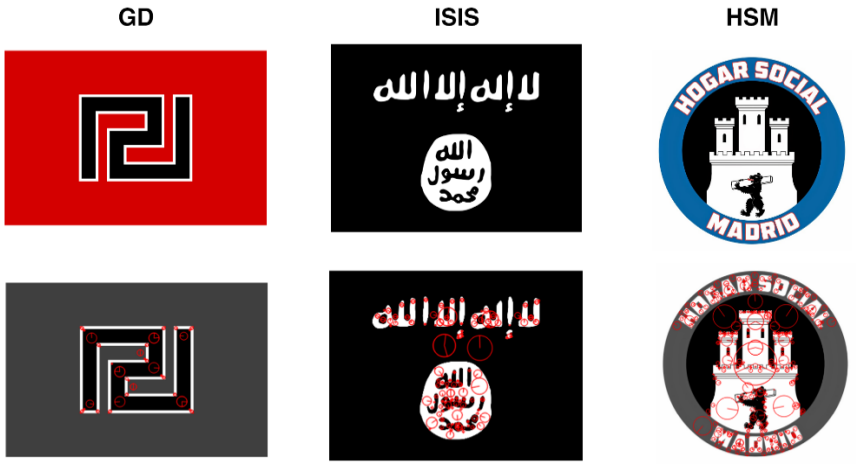

Fig. 2. Reference images for Golden Dawn, Islamic State and Hogar Social Madrid (left to right), and key points as detected by the SIFT algorithm (second row)

After selecting the reference images, we collected a database with real-world occurrences of the reference images. In total, we collected 200 images distributed over 26 GD images, 57 ISIS images and 57 HSM images. Additionally, 60 images were collected containing no occurrence of the reference images. These 60 images were selected from a variety of topics to account for the inherent variability of online images. Figure 3 shows some sample images from each category. To evaluate the different feature descriptors discussed in the previous section, we computed the descriptors of each of the reference images and, then, used them to try to retrieve all the images with the same iconography from the collected dataset. In each experiment, the images containing iconography of the same group as the reference image were considered as positive, while images containing iconography of other groups were considered as negative together with the images containing

no iconography at all. For each experiment, we computed the ROC curve (figures 4, 5 and 6) and the corresponding AUCs (table 1).

Fig. 3. Representative images from each category in the collected dataset

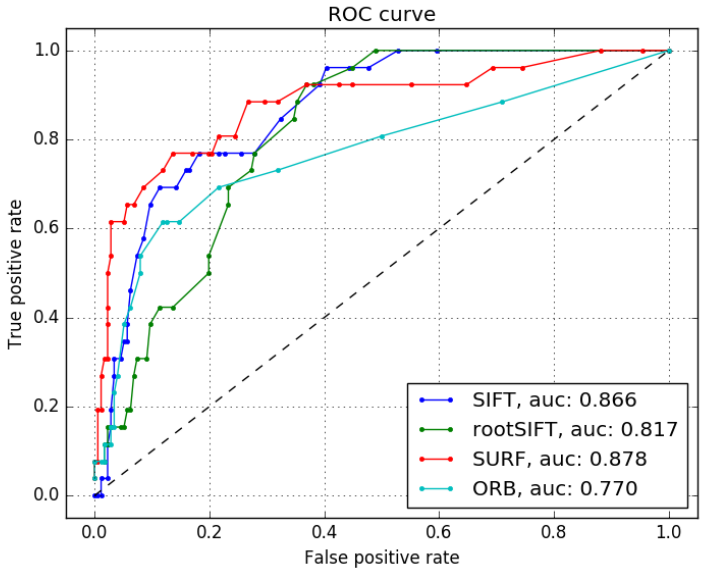

Fig. 4. ROC curve comparing different feature descriptors, retrieving GD images

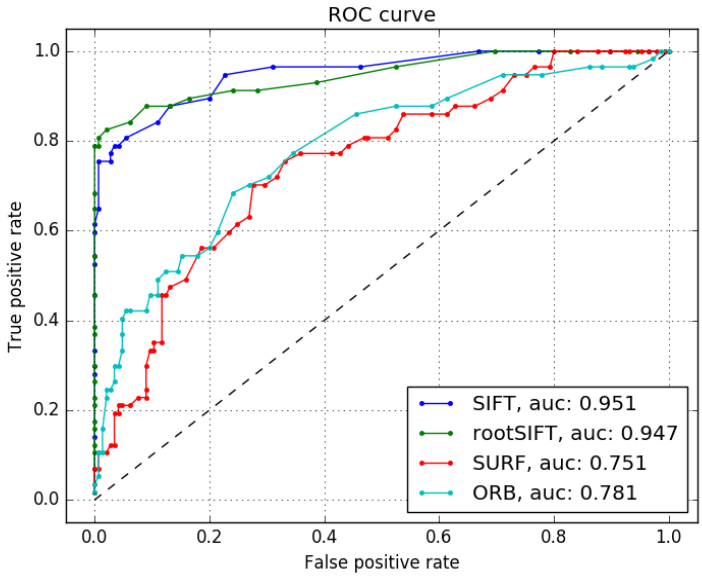

Fig. 5. ROC curve comparing different feature descriptors, retrieving ISIS images

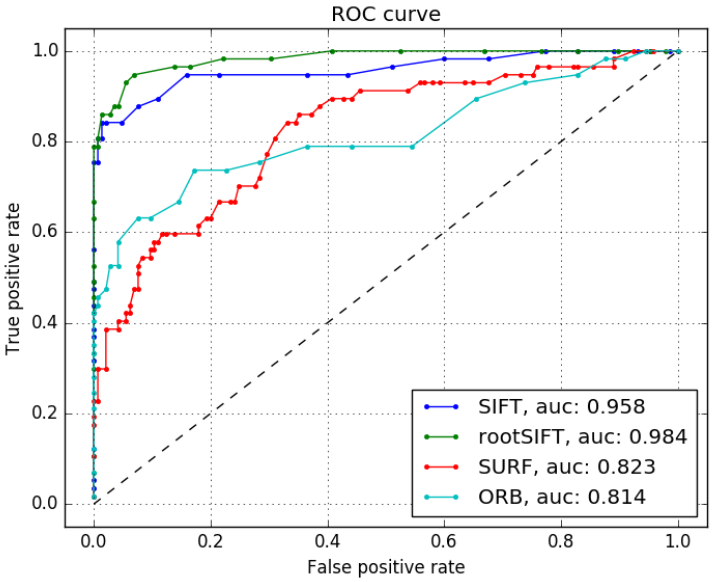

Fig. 6. ROC curve comparing different feature descriptors, retrieving HSM images

ROC AUCs			
Group	**SIFT**	**Root-SIFT**	**SURF**
GD	0.866	0.817	0.878
ISIS	0.951	0.947	0.751
HSM	0.958	0.984	0.823

Table 1: Areas under the ROC curve (ROC AUC) for the different experiments

4 Discussion and future work

The experimental results presented in the previous section evidence the potential of local feature descriptors in the task of detecting radical iconography in unconstrained real-world images. Looking at figures 4, 5 and 6, we can see that the retrieval accuracy was significantly greater for HSM and ISIS than for GD. We believe this is a consequence of the low visual complexity of the iconography associated with GD. The absence of distinctive visual features makes it harder for all the descriptors evaluated to retrieve the correct images. In addition, our results show that, except for the case of GD, SIFT greatly outperformed SURF and ORB descriptors. This suggests that SIFT might be the best option to detect iconography of extremist groups in online images, as long as its computational cost is admissible in the specific application case. Regarding the comparison between SIFT and Root-SIFT, no conclusive results were found, and further experimentation is required to determine whether Root-SIFT outperforms SIFT in this context.

In this study, we matched the feature descriptors of a reference image with those of target images only to count the number of matches (according to Lowe's ratio test [8]). Then, normalizing this number of matches we obtained a probability for the presence of the reference image in the target image. Nevertheless, it is possible to use the matches to compute the estimated location of the reference image in the target image. This might be useful to provide a justification of the recommendations of the platform to final users, increasing the explainability of the system. Figure 7 shows an example of this type of location estimation, which can be achieved with standard computer vision tools such as OpenCV [5].

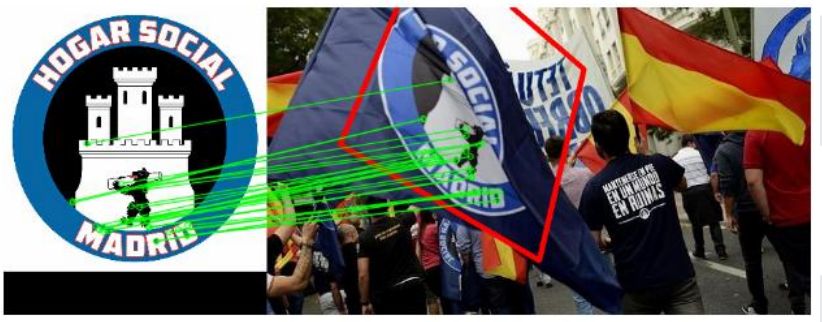

Fig. 7. Example of successful detection of iconography in real-world images, additionally showing the estimated location of the reference image in the target image (HSM).

During our experiments, we have also identified some possible drawbacks of the described approach. For instance, some of the images containing the flag of ISIS were actually of Iraqi soldiers holding the flag upside-down as a sign of victory after recapturing Mosul. In this case, users sharing this image are unlikely to be ISIS supporters. However, as the image descriptors applied are rotation invariant, this image is detected as a positive match by our approach (see figure 8). This type of particularities require a certain degree of expert knowledge to be decided and shall probably be left to the judgement of the human expert, whose work is being assisted by the automated system.

Fig. 8. ISIS flag is detected on an image of Iraqi soldiers celebrating the recapture of Mosul

As for future research lines, we are yet to define how the iconography detection approach presented in this paper can be combined with the risk function defined in equation 3. A simple approach would increase the value of the risk function for a given user if they have shared or used as profile image a picture containing radical iconography. However, the integration of these approaches (i.e., interaction-based risk estimation and radical iconography detection) might be performed in a more sophisticated manner. We also intend to explore the role of reciprocity in social interactions for radicalization risk assessment. For instance, mentions from confirmed radical users to profiles which never answer them and have a significantly higher number of followers might not be indicative of possible ongoing radicalization processes.

References

[1] Agarwal, S., & Sureka, A. (2015). Using knn and svm based one-class classifier for detecting online radicalization on twitter. In International Conference on Distributed Computing and Internet Technology (pp. 431-442). Springer, Cham.

[2] Arandjelović, R., & Zisserman, A. (2012, June). Three things everyone should know to improve object retrieval. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 2911-2918). IEEE.

[3] Bay, H., Tuytelaars, T., & Van Gool, L. (2006, May). Surf: Speeded up robust features. In European conference on computer vision (pp. 404-417). Springer, Berlin, Heidelberg.

[4] Bermingham, A., Conway, M., McInerney, L., O'Hare, N., & Smeaton, A. F. (2009). Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In Social Network Analysis and Mining, 2009.

[5] Bradski, G., & Kaehler, A. (2008). Learning OpenCV: Computer vision with the OpenCV library. " O'Reilly Media, Inc.".

[6] López-Sánchez, D. and Revuelta, J. and De la Prieta, F. and Gil-González, A. B. and Dang, C. (2016). Twitter User Clustering Based on Their Preferences and the Louvain Algorithm. In Trends in Practical Applications of Scalable Multi-Agent Systems, the PAAMS Collection (pp. 349-356). Springer, Cham.

[7] López-Sánchez D., Revuelta J., de la Prieta F., Corchado J.M. (2018). Towards the Automatic Identification and Monitoring of Radicalization Activities in Twitter. In: Uden L., Hadzima B., Ting IH. (eds) Knowledge Management in Organizations. KMO 2018. Communications in Computer and Information Science, vol 877. Springer, Cham

[8] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

[9] Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011, November). ORB: An efficient alternative to SIFT or SURF. In Computer Vision (ICCV), 2011 IEEE international conference on (pp. 2564-2571). IEEE.

[10] Thelwall, M. (2017). The Heart and soul of the web? Sentiment strength detection in the social web with SentiStrength. In Cyberemotions (pp. 119-134). Springer, Cham.