Filtering Local Features for Logo Detection and Localization in Sports Videos

Daniel Manger, Markus Müller,
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB)
Karlsruhe, Germany

Markus Kubietziel
ValuMedia GmbH
Stuttgart, Germany

***Abstract*— This paper presents a system for the detection and localization of multiple instances of trademark logos in sports videos. It is based on SIFT features and considers the local geometry of neighboring features in order to differentiate between different logos with ambiguous local features such as text-based logos. In contrast to other approaches, we do not rely on a training phase and therefore no labeled data with annotated or absent logos is needed. The focus of the detection approach is on images of sports videos which suffer from compression artifacts, motion blur, small object sizes, occlusion and several other artifacts. Results are presented on video images of a soccer game containing logos on different advertising media.**

Keywords*— *logo detection, sports videos, trademark, advertising, matching, retrieval

I. INTRODUCTION

Every year, many companies spend a considerable amount of their advertisement budget on sponsorships and sports marketing. A large part of it is for placing logos, trademarks or company names on banners, billboards, shirts, bottles and such which are assumed to be visible in TV and web broadcasts. For evaluating the return on the sponsorship investment, it is essential to measure the actual visibility of the advertising media in the consumers's view. To this end, many sports marketing firms offer services to verify the media visibility. Often, this is performed manually by watching the relevant sports videos and manually timing the appearance of the advertising media. Since this is a tedious, expensive and not always replicable task, (semi-)automatic systems have attracted the attention of researchers in recent years.

II. PREVIOUS WORK

Early approaches for logo detection used low-level features such as edges detected by canny filters and morphological operations [5] or classifiers [10] to generate logo detections. [12] used color co-occurrence features which are however not always discriminative enough, especially for different logos with similar colors. As in many other applications of computer vision, the introduction of local features such as SIFT [11] has improved the performance due to its robustness to certain transformations and to noise or background clutter. For example, [2] and [3] use SIFT features and determine matches between logo features and video features using Lowe's distance ratio of the closest neighbor to the second-closest neighbor in descriptor space [11]. This turned out to be a robust and easy choice assuming that unique features predominate. In the context of trademarks in sports videos, however, there are often multiple instances of the same logo next to each other on a banner which renders the ratio check ineffective. Furthermore, even individual logos consisting of simple letters tend to have ambiguous features at the corners of the individual characters. Similarly, considering just the best matching feature [7] and discarding matches above a global distance threshold [16] doesn't work either for multiple instances. In other words, when using local features, the position of the features within the logos and in relation to each other is very important. Incorporating this spatial configuration of the local features for logo detection has been addressed by several papers. Typically, features are quantized to visual words and a model of the spatial configuration is learnt in a supervised manner using a training dataset of logo appearances and background images. E.g., in [9], spatial pyramid mining is introduced to build a model, whereas [15] includes the layout of triples of visual words into visual

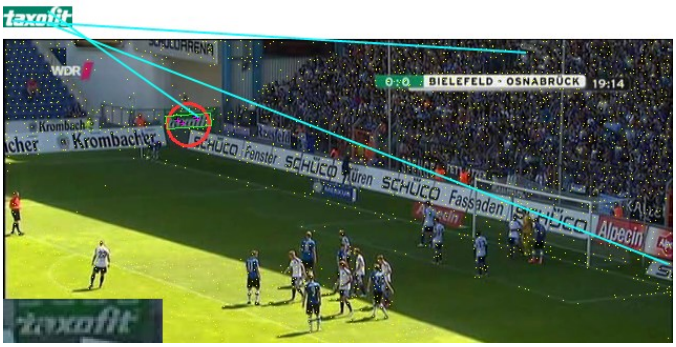

Fig. 1: Why local geometry matters: in this image containing 3,493 SIFT features (yellow pixels), directly matching the logo features using the well-known distance ratio only yields three matches (cyan lines) of which only one is correct. Taking into account the local geometry of neighboring features, the best match (feature with the red pixel, surrounded by the red circle for better visibility) has four locally consistent features (magenta crosses) in the neighborhood enabling a correct detection (green rectangle). In the lower left, the logo appearance is shown enlarged.

signatures for indexing. [6] describes the spatial relationship of two features with an invariant representation and subsequently adopts data mining techniques to find visual patterns. [8] locally groups feature triples using multi-scale Delaunay triangulation for large-scale logo recognition. [14] proposes bundling features in a spatial neighborhood and indexing their visual word occurrences by min-hashing. In [13], the burstiness problem for logo retrieval is addressed, i.e. the fact that logo features are frequently found in random background structures in a similar configuration. Therefore, a statistical model is learnt for the distribution of incorrect detections output leading to a noticeable boost in performance.

Apart from the fact, that these approaches need labeled training databases, most of them focus on the retrieval aspect for large sets of different logos or on learning a model to retrieve several variants of a trademark, i.e. aiming at the category-level rather than the instance-level. Furthermore, the used datasets consist of photos which are different to frames of a sports video in terms of quality. In sports videos, due to several artifacts, even when using unquantized features and exhaustive nearest neighbor search, often 90% of the logo features cannot be found in the video images containing instances of the logo. In these cases, finding pairs or triples of quantized features (visual words) which are furthermore consistent with the learnt model can end up in a game of patience.

We propose an iterative local feature-based detection and localization approach for which no training data is needed. The local geometry of features is not learnt in a model but considered in the matching step instead. Subsequently, several filter steps are performed in order to reject erroneous detection hypotheses.

Section 3 describes the matching process for determining the first and most plausible detection hypothesis in an image. In Section 4, both the filtering steps are described to reject false positives and the iterative process to arrive at the next hypothesis. Finally, Section 5 deals with experiments followed by a conclusion.

III. Detecting Hypotheses

Given a logo image and a test image, the aim of this step is to find a detection hypothesis based on local features incorporating both descriptor information and geometry of the features. As stated in section 2, when matching the two sets of local features of the involved images, the distance ratio check is a contrast to finding multiple instances because the matches might spread to different logo instances. We therefore keep up to k nearest matches above a global threshold τ. Next, for each feature in both images, subsequently referred to as central feature, a list of its relevant local neighbor features is calculated: depending on the central feature scale σ_c, a radius $\alpha\sigma_c$ is defined for the neighborhood. α is chosen such that in a logo with text, the largest (text-related) features define a neighborhood which is approximately half the size of the logo. Features in this neighborhood are added to the list if their scale ratio with respect to the central feature satisfies $\frac{1}{3} < \frac{\sigma_n}{\sigma_c} < 3$ since typically, matching features relating to the same object mostly appear on nearby scales.

From all feature matches between the logo and test images and the respective two neighbor feature lists of each involved two features, we now exhaustively search for the most plausible match in terms of geometrical consistence of the neighboring feature sets. Considering one match, let $f_l = (\boldsymbol{x}_l, \theta_l, \sigma_l)$ be the logo feature with its image coordinate $\boldsymbol{x}_l$, dominant orientation θ_l and scale σ_l and $f_t = (\boldsymbol{x}_t, \theta_t, \sigma_t)$ the corresponding test feature of one match. Furthermore, we denote by $f_l^n = (\boldsymbol{x}_l^n, \theta_l^n, \sigma_l^n)$ one of the logo features of the neighbor feature list of f_l and $f_t^n = (\boldsymbol{x}_t^n, \theta_t^n, \sigma_t^n)$ one of the logo features of the neighbor feature list of f_t, respectively. For legibility, we omit indices to indicate the individual matching features of the two images and the individual neighbor features of one of the matching features. For every match, from the two sets of neighboring features we regard those feature pairs which matched in descriptor space and pass the following filtering steps to ensure locally consistent geometry:

$$\left|\frac{\sigma_l}{\sigma_t} - \frac{\sigma_l^n}{\sigma_t^n}\right| < 0.5$$

and

$$\tau[\varphi(\theta_t - \theta_l) - \varphi(\theta_t^n - \theta_l^n)] < \frac{45°}{360°} 2\pi,$$

where $\varphi(\cdot)$ maps the orientation differences to be in the interval $[-\pi \dots \pi)$ by adding or subtracting 2π, and $\tau[\cdot]$ ensures its argument to be in the interval $[0 \dots \pi)$ by subtracting if from 2π if necessary. Finally, we check if the distances from f_l to f_l^n are consistent with those from f_t to f_t^n taking into account the involved feature scales:

$$\left|\frac{\sigma_l}{\sigma_t} - \frac{\|\boldsymbol{x}_l - \boldsymbol{x}_l^n\|}{\|\boldsymbol{x}_t - \boldsymbol{x}_t^n\|}\right| < 0.5$$

The number of all neighboring features of a match which pass all three filtering steps are counted and a hypothesis is created for each match having at least three of them.

IV. Filtering Hypotheses

Most advertising media are logos on flat banners with a well-known digital template. Thus, at first glance, logo detection could be considered as a near-duplicate retrieval or detection task. However, several artifacts such as occlusion by players, small object resolution, compression artifacts, motion blurriness, etc. renders it a more complex task and in fact, the matching of features in descriptor space has to be carried out with a quite tolerant global threshold. Although having taken into account a lot of information in the hypothesis detection step, there are still false positives which coincidentally agree in terms of local geometry. To filter out as much of these false positive hypotheses, we apply the following steps:

The scales and orientations of the two features of the central match define the similarity transform which yields the expected position of the logo in the test image. The false positive detections often have matching features lying on a line in the test image or spreading over a wide part of the image (larger than the estimated logo dimensions) although

the involved features showed locally consistent orientations and scales (see Section 3). We therefore first filter hypotheses based on the ratio of the area of the convex hull of the involved features in the test image to the area of the logo hypothesis. For the erroneous features on a line, this ratio is close to zero whereas for the spreading features, it is often much larger than one.

Second, to check the positions of the involved features in the test image, we project each logo feature position according to its feature match (scale and orientation difference) to the location in the test image. For correct matches, this location would be close to the actual location of the corresponding matching feature in the test image whereas for incorrect matches, this position often is somewhere else inside the detection hypothesis or even outside. Averaging all those distances between the projected and real feature position in the test image and normalizing it by the longer side of the detection hypothesis box to preserve scale invariance, we apply a threshold and filter out many of the false positives occurring from other logos with similar letters at sometimes even similar positions within the logo. Last, we only allow detection hypotheses which are rotated up to +/- 90° compared to the logo template since typically, logos do not appear upside down.

When a hypothesis passed the three checks above, a new detection is generated. Subsequently, all features inside the detected logo rectangle are removed from the test image. If the hypothesis did not pass the checks, a false positive is assumed and only the involved features which led to the detection hypothesis are removed. Of course, the list of matching features and the relevant neighborhood features are updated accordingly. Finally, the process of searching the "next best" hypothesis (see Section 3) is performed again, so that, iteratively, all instances of the logo might be detected. The process is stopped when the geometric plausibility score gets too low.

V. Experiments

For the evaluation of logo detection systems, most authors use the Flickr-Logos dataset [15] or the BelgaLogos dataset [7]. However, these do not accurately focus on the scenario our system is intended for. The former dataset contains images for the category-level detection task with quite different variants of trademarks. Moreover, most images are photographs which is true for the latter dataset as well. Therefore, we use an own dataset recorded from television broadcast. In a complete sequence of a soccer game with a resolution of 640x360 pixels we extracted one frame every second yielding some 7,300 frames and annotated the occurrences of one trademark logo which is visible in 168 frames.

In all frames, we extract features on DOG interest points and use the RootSIFT [1] variant for descriptors since DOG interest points are a common choice for images with noise and blur. To detect both variants of the logo (in white font on dark backgrounds and vice versa), we use one image of both variants as logo template. Furthermore, the logos appear in the images in various sizes. Although the features themselves are scale-invariant, too large scale differences lead to different features encoding different levels of details. In small details such as the ® character, edges are becoming blurred if the logo gets too small. We take this into account by generating six versions of each logo template with different sizes (376 to 42 pixels in width). Both template augmentations (inverting colors and different sizes) are integrated in the hypothesis generation step described in Section 3, i.e., in each iteration, the most appropriate version might be chosen out of the 12 logo templates (Fig. 2).

Using a bounding box overlap of 50% for determining true and false positives and with all parameters chosen empirically to maximize recall, the system outputs 130 true positives and 436 false positives for the 7,323 frames. For a productive system this still can be of use since, in the dataset, we annotated even very small occurrences of the logo which might not catch the attention of the viewer. Furthermore, the results are detections based on single image frames not considering the possibilities of tracking or temporal filters to reject scattered false positive detections often arising in the visitor area in random frames. Figure 3 shows some examples of true positives, false positives and false negatives.

VI. Conclusion

We outlined a logo detection and localization system. Without a training phase, multiple instances of logos can be found in sports videos. Future work will contain more experiments and could deal with an approach to explore the rejected hypotheses in an unsupervised way in order to automatically determine other logos with similar feature configurations and to explicitly detect them for removal in a second run of the detection workflow.

References

[1] Arandjelovic, Relja, and Andrew Zisserman. "Three things everyone should know to improve object retrieval." Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.

[2] Bagdanov, Andrew D., et al. "Trademark matching and retrieval in sports video databases." Proceedings of the international workshop on Workshop on multimedia information retrieval. ACM, 2007.

[3] Ballan, Lamberto, et al. "Automatic trademark detection and recognition in sport videos." Multimedia and Expo, 2008 IEEE International Conference on. IEEE, 2008.

[4] Ballan, Lamberto, Marco Bertini, and Arjun Jain. "A system for automatic detection and recognition of advertising trademarks in sports videos." Proceedings of the 16th ACM international conference on Multimedia. ACM, 2008.

[5] Chattopadhyay, T., and Aniruddha Sinha. "Recognition of trademarks from sports videos for channel hyperlinking in consumer end." Consumer Electronics, 2009. ISCE'09. IEEE 13th International Symposium on. IEEE, 2009.

[6] Chu, Wei-Ta, and Tsung-Che Lin. "Logo recognition and localization in real-world images by using visual patterns." Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012.

[7] Joly, Alexis, and Olivier Buisson. "Logo retrieval with a contrario visual query expansion." Proceedings of the 17th ACM international conference on Multimedia. ACM, 2009.
[8] Kalantidis, Yannis, et al. "Scalable triangulation-based logo recognition." Proceedings of the 1st ACM International Conference on Multimedia Retrieval. ACM, 2011.
[9] Kleban, Jim, Xing Xie, and Wei-Ying Ma. "Spatial pyramid mining for logo detection in natural scenes." Multimedia and Expo, 2008 IEEE International Conference on. IEEE, 2008.
[10] Kovar, Bohumil, and Alan Hanjalic. "Logo detection and classification in a sport video: video indexing for sponsorship revenue control." Electronic Imaging 2002. International Society for Optics and Photonics, 2001.
[11] Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.
[12] Phan, Raymond, John Chia, and Dimitrios Androutsos. "Colour logo and trademark detection in unconstrained images using colour edge gradient co-occurrence histograms." Electrical and Computer Engineering, 2008. CCECE 2008. Canadian Conference on. IEEE, 2008.
[13] Revaud, Jerome, Matthijs Douze, and Cordelia Schmid. "Correlation-based burstiness for logo retrieval." Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012.
[14] Romberg, Stefan, and Rainer Lienhart. "Bundle min-hashing for logo recognition." Proceedings of the 3rd ACM conference on International conference on multimedia retrieval. ACM, 2013.
[15] Romberg, Stefan, et al. "Scalable logo recognition in real-world images." Proceedings of the 1st ACM International Conference on Multimedia Retrieval. ACM, 2011.
[16] Sanyal, Subhajit, and Srinivasan H. Sengamedu. "Logoseeker: a system for detecting and matching logos in natural images." Proceedings of the 15th international conference on Multimedia. ACM, 2007.

Fig. 2: All 12 logo templates used for the detection.

Fig. 3: Examples of detection hypotheses. Green: true positives, red: false positives. The red crosses denote the positions of the central match features, the smaller yellow crosses mark the positions of the neighboring features. The green lines at those features indicate the distances to the projected position.