computer graphics & vision

Inffeldgasse 16/2, A-8010 Graz

www.icg.tu-graz.ac.at

TUG

Planar Trademark and Logo Retrieval

Martin Köstinger, Peter M. Roth, and Horst Bischof

Inst. for Computer Graphics and Vision
Graz University of Technology, Austria

Technical Report
ICG–TR–10/01
Graz, February 15, 2010

Abstract

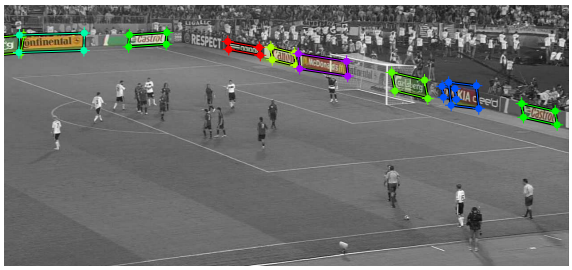
Sport advertising has become an important business increasingly raising the interest of an efficient analysis. To reduce the manual workload, in this work we present an automatic specific trademark and logo recognition system overcoming typical problems of existing (mostly SIFT-based) approaches. In particular, we need to cope with relatively small or correlated trademarks, severe background clutter, and huge perspective variations. This is realized by introducing a concept to increase the perspective invariance, a sophisticated verification, a guided matching phase that is able to deal with a vast number of outliers, and the use of an additional complementary interest region detector with multi-resolution shape description. To show the benefits of the approach, we demonstrate it for a representative real-world test set consisting of images of the EURO 2008 final game. The results clearly show that using the proposed method existing approaches can be outperformed in terms of accuracy and recall.

Keywords: *logo retrival, trademark retrival, SIFT*

1 Introduction

Sport advertising evolved to a multi-billion industry with the clear and precise focus to influence consumers positively on products or services. Thus, a major goal of the industry is to rate the effectiveness of advertisements. In the case of sport broadcasts, as application domain for trademark and logo retrieval, a suitable measure would be the visible time. Nevertheless, automatic visibility analysis of on-site means of promotion as billboards is not straight forward. Hence, even if laborious, the required data is still acquired mostly by manual annotation. To overcome these circumstances, computer vision has the potential to serve as key technology to automate and speed up this process. In general, computer vision based trademark and logo recognition approaches can be coarsely categorized into three main groups: (a) sport specific billboard detection and recognition, (b) virtual advertisements, and (c) unconstrained billboard, trademark, logo detection and/or recognition. Approaches motivated by intellectual property protection concerns can be delimited. They concern mainly the perceptual similarity of logos [3, 6, 8, 11, 22].

(a) Baseline method: Retrieval approach based on SIFT [15].



(b) Proposed approach.

Figure 1: Trademark retrieval: (a) Approaches based on plain SIFT [15] can not deal sufficiently with problems such as small trademark occurrences under severe perspective changes; (b) In contrast the proposed method handles these problems considerable better.

Billboard detection and recognition approaches, *e.g.*, [13,23] are restricted on a very specific application domain, in form of the sports ground. In [13] the basic idea is to detect the delimiting lines of billboards next to the green with a Hough transform based on Canny edges. Hereby, also knowledge about the color distribution of the sports ground and the contrast of the surrounding billboards is assumed. Moreover, adjacent billboards are delimited based on the background color as splitting criterion, which is not applicable in all cases. In addition, there are several limitations if there are billboards on both sides of the sports ground and if they are partially occluded.

Similar to [13], in [23] the goal is to detect and to identify billboards in soccer videos. Also relying critically on analyzing detected edges. Based on a mixture of global and local features the individual billboard is subdivided into several regions in which features such as mean, variance, hue, and intensity are extracted and stacked into a descriptor to classify. The location of identified billboards is used to search for further billboards in near proximity.

In the case of virtual advertisements the goal is to overlay a seamless replacement for existing advertisements or potential advertisement regions. For that purpose, the pose of the to be replaced billboard, trademark, or region has to be estimated in the image. A viable solution is given in [18], which examines the target position of the virtual advertisement by interest points and color filtering of these and geometric hashing. This method implicitly identifies the overlayed billboards or regions unconstrained of the position, unless enough interest points are detected to distinguish it. However, the authors only provide a few example images of large perpendicular billboards; no information about the performance and accuracy is given.

In contrast the work of [5] is concerned with trademark recognition by matching the text only. This is realized by first searching for homogeneously colored regions that are surrounded by large color differences. Then along the longest foreground transition a descriptor is extracted that provides only binary foreground or background flags for each line segment. This is based on the observation that the line position is affine invariant and thus the calculated descriptor can then be used to classify the logo. However, the presented results show bad recall and precision.

A more sophisticated method was proposed in [10], where potential trademarks or billboards are identified with a probability model based on chrominance histograms. However, to reduce illumination effects the color model for each target has to be acquired under the actual illumination conditions. With increasing database size the performance drops severely, which limits application in practice.

Indisputably, the recent success of specific object recognition based on local features emphasis the idea to accomplish the task of trademark and logo retrieval. Local features are per definition robust to occlusions and clutter. In addition, no prior segmentation is needed. Furthermore, even for small objects an adequate number of distinctive features can be generated. Hence, Scale Invariant Feature Transform (SIFT) [15] based approaches (see Fig. 1(a)) such as [1] are very promising; although this particular approach lacks in terms of pose estimation, since the location of the trademark is only approximated. Thus, a top down back-projection step, e.g., for guided matching of small trademarks, delivering only few matches, is not possible. Furthermore, the normalized match threshold used does not take into account the feature density in the region of interest in the query image nor in the projected training image.

In this paper, we address the problem of planar trademark and logo recognition unrestricted of a specific setting or sport. Although trademark recognition was previously faced none of the proposed strategies is proper for our purpose. In particular, our approach (see Fig. 1(b)) uses the advantages of a local feature-based approach and tackles the key problems arising in the problem domain.

As first contribution the problem of little trademarks with only few potential matches and a high outlier rate is covered by an accurate pose estimation and guided matching. Also an additional complementary interest region detector is incorporated to induce more robustness. Indisputably, dealing with huge view point changes is absolutely essential for the approach. To introduce more invariance we propose to pre-train the system on synthesized views of the trademarks. To show the benefits of the proposed method, we demonstrate it on real world data in particular on a representative test set of the EURO 2009 final game, obtaining excellent results.

The remaining of this paper is organized as follows. In Section 2 we discuss the currently arising problems and motivate the proposed approach. The building blocks of the layered retrieval approach are described and discussed in Sections 3-5. An overview of the whole approach is given in Section 6. Detailed evaluations of the proposed approach comparing to the baseline are given in Section 7. Finally, we summarize and conclude the paper in Section 8.

2 Motivation

Motivated by the suitability of local object recognition methods for the task of planar trademark or logo retrieval, we follow the basic structure, outlined in Figure 2. Since a retrieval approach based on Lowe's SIFT [15] is a reasonable basis, we use it as starting point for further investigations. Nevertheless, a plain SIFT-based approach has several drawbacks. Among others, mutual blocking trademarks and problems with the perspective invariance.

(a) IR detection: Model image.

(b) IR detection: Scene image.

(c) IR description: Model image.

(d) IR description: Scene image.

(e) IR Matching.

Figure 2: Basic flow of object recognition with local features: Interest regions (IRs) are extracted (a)-(b) and described (c)-(d). Interest region correspondences are formed (e). The green lines indicate correct match correspondences, estimated pose shown in dashed green.

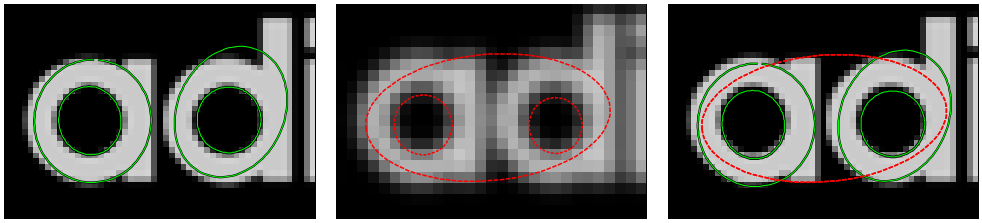
Little or low textured trademarks are problematic, since they only deliver few potential matches. Therefore, using and incorporating an additional complementary interest region detector might be beneficial. In particular, if the characteristics of trademarks can be described by clear outlines with relatively homogeneous texture Maximally Stable Extremal Regions (MSER) [16] would provide excellent results. MSER works particularly well in this configuration and yields also good performance in terms of repeatability under geometric transformations. Even though the number of detected regions is relatively small compared to the Difference-of-Gaussian (DoG) detector. The topic of interest region detection and description is covered in more detail in Section 3.

In typical application scenarios the trademarks are subject to severe perspective changes. Hence, for a trademark recognition system it is desirable to cope with these variations. Typically, up to some extent the detector / descriptor combination is insensitive, but often it is impossible to get reliable match correspondences. For instance, Lepetit et al. [12] use a small number of training images to synthesize many new affine views of each detected interest region. The result is an artificial view set for each patch which should reflect the set of all of its possible appearances under different viewing conditions. In this context, due to the exhaustive description, efficient nearest neighbor search is of interest. Further details on our solution for this problem are given in Section 4.

An ideal matching strategy provides high recall and precision. Nevertheless, there is typically a trade-off between precision and recall. Thus, our matching strategy concentrates on keeping the recall as high as possible, however still preserving the precision. Experiments showed that in the case of trademark and logo retrieval the recall of the standard SIFT matching approach noticeable decreases on increasing the number of trained trademarks. Basically it is assumed that this can be traced back on the similarity of the trained trademarks, which are mostly homogeneous and therefore not that distinctive. Moreover, they also contain same letters with similar or even equal fonts. The dropping recall leads us to an adaptation of the matching scheme, that is also suited to deal with a large proportion of false matches. Furthermore, we suggest to merge the trademarks of the same brand into a common representation, which should abolish the problem of too correlated logos that block each other in the standard matching scheme. Since the problem of small trademarks also considers the matching, we propose to use a guided matching in conjunction with an accurate pose estimation to deliver more match candidates. The basic structure of our matching strategy is to form region to region correspondences that are clustered into pose hypotheses for particular training images which are finally verified.

3 Interest Region Detection and Description

In a local feature-based approach basically each additional feature type contributes robustness. For instance, MSER as complementary detector to DoG is assumed to be well suited for that purpose. In our approach, we apply a simple extension to MSER, referred to *'Multi-Resolution MSER'* [9], which is based on the idea that the extremal regions are only detected in a single image resolution. However, if an image is viewed from increasing distance many details in the image disappear and different region boundaries can emerge, especially with additional view point changes. Therefore, a scale pyramid is constructed by Gaussian blurring and sub-sampling from which the MSERs are extracted separately at each image resolution. Subsequently duplicate MSERs are removed by eliminating the fine scale MSERs with matching them in terms of location and size with the next coarser scale. This is illustrated in Figure 3.



(a) Original patch. (b) Sub-sampled. (c) Combined.

Figure 3: Multi-Resolution MSER: (a) Original image patch with inpainted MSER covariance ellipses; (b) Due to sub-sampling and Gaussian blurring a new outer region boundary arises; (c) Merged representation.

As we want to incorporate the MSER detector in the existing DoG / SIFT concept further problems arise. The MSER detector delivers shape of the region, but neither orientation information nor a scale measure. A viable solution to overcome these problems is to use a normalized coordinate system, which suggests the use of local affine frames (LAFs). This basically enables an affine invariant normalization of the region. Matas et al. [17] propose two possible construction methods for LAFs. On the one hand stable bi-tangents and on the other hand region normalization by the covariance matrix. Both methods have been used for wide-baseline stereo [17] and object recognition tasks [4, 20], respectively.

In our approach, we use the LAF covariance construction, as illustrated in Figure 4. The remaining rotation ambiguity is resolved over the normalized contour distances between the center of gravity and the contour pixels. We

encountered that an assignment by gradient orientation histogram can be error prone if minor detector errors occur. Strong gradients, e.g., a part of an adjacent letter, are likely to alter the result.

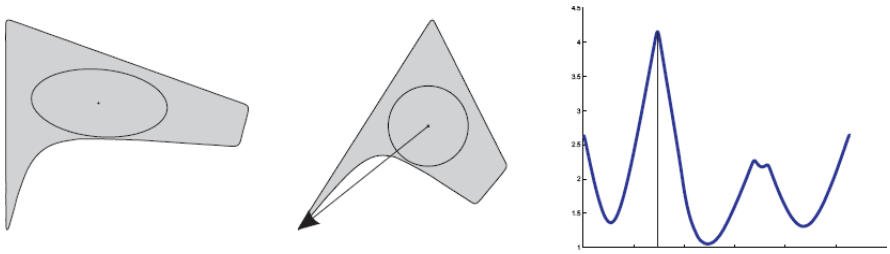


Figure 4: Local affine frame covariance construction. The detected region is transformed into the affine normalized form by the inverse square root of the covariance matrix. To solve the rotation ambiguity the normalized contour distance is considered (adapted from [20]).

The remaining step after region normalization and orientation assignment is the actual description. One solution is to compute the SIFT descriptor on the texture of the image patch. Another possibility is to calculate the descriptor on the actual shape of the detected region [9]. Hereby, the descriptor is calculated on the binarized region of the affine normalized patch. To match the shape descriptors it is suggested to use the χ^2 distance measure in favor of the Euclidean distance [9]. One major advantage of the shape descriptor is that it only depends on the correct detection of the region. Thus, background clutter has only little influence. Moreover, the shape descriptor is more robust in terms of illumination changes and is therefore proposed as reasonable extension to strengthen the performance under difficult illumination conditions.

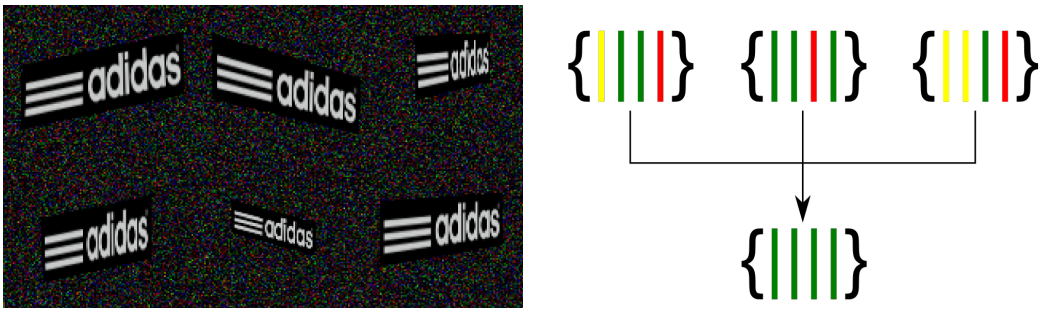
4 View Set based Stable Model Generation

It is desirable for a trademark recognition system to cope with severe perspective changes. Therefore, we want to generate an arbitrary number of artificial views of each trademark to pre-train our system, similar to [12, 21]. Our approach, however, differs in several ways. To construct the synthetic view set, we use labeled homographies. This is not as exhaustive as regular sampling of the parameter space and can either be generic or task specific. Furthermore, we transform the whole image; not each individual patch. Thus, for each view the interest regions are detected and described. In this way, the localization errors of the detector are bypassed. We propose two methods

to combine the different representations of the view set into a common one, which we refer as "stable model". First, we encourage the idea to select descriptors equally privileging each view under a given criterion. Later on, a data structure is proposed that is capable to deal with all descriptors of all views, also providing efficient retrieval.

4.1 Selection of Stable Descriptors

To build a compact representation of the view set we conduct a feature stability assessment. The aim is to rank and select only the most stable descriptors, e.g., correctly matched and geometric consistent over as many views. Obviously, most of the regions in a particular view have at least one corresponding region in each similar view. Therefore, the selection criterion ensures that the selected descriptors are not too similar and that the resulting representation is compact although discriminative. A voting model preserves that each view is equally privileged. Figure 5 illustrates this selection process.



(a) View set. (b) Descriptor selection.

Figure 5: Stable descriptor selection -well suited descriptors are displayed in green.

As first selection criterion, we propose to select the regions based on detector errors that result in differing descriptors. The second criterion enforces a minimum distance in feature space. Thus, it rejects interest regions which are below a certain distance threshold to previously selected ones, without explicitly insisting on a different location, scale or orientation.

4.2 k-Means View Tree

The view set of each trademark provides a rich, although exhaustive, description of possible appearances, containing much redundancy. On the one hand this description promises reasonable matches even under severe perspective

changes. On the other hand an efficient retrieval strategy is required. Due to the vast alternatives of corresponding, geometrically consistent, descriptors the nearest neighbor retrieval can be coarse. Even a simple nearest neighbor (NN) search would provide reasonable matches. Moreover, the NN retrieval can be approximate and even coarser, as, e.g., in the *'best bin first'* K-D Tree [2] and therefore much faster. For instance Nister uses in his large scale object recognition system [19] a hierarchically quantized k-means tree to provide fast access to the leaf nodes, which store scoring information. In our k-means tree (Fig. 6) the actual descriptors are indexed in the leaf nodes, not the scoring information. A query in the k-means view tree is rather efficient, since in our case the tree has a depth of 3 - 4.

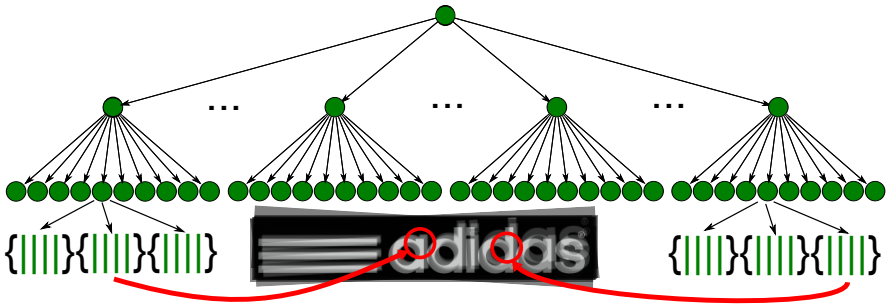


Figure 6: K-means view tree. Nearest neighbor assignment is carried out by traversing down along the path of the nearest tree nodes. At the resulting leaf node the nearest descriptor is determined.

To compact the tree, it is proposed to prune on the basis of the geometric consistency of the interest regions. Thus, if the interest regions in a particular node are geometric consistent, the node itself lays out the interest region descriptor on basis of the centroid. The meta data such as spatial coordinates, scale and orientation is obtained clustered of the corresponding regions. The pruning does not affect the descriptor matching, since before pruning each query descriptor had to pass the node too. Starting at the leaf nodes this procedure can be applied as recursive ascend on the tree eliminating much redundancy. Obviously the k-means view tree can be extended to contain many models. Nevertheless, this will lead in the best case to similar results and therefore it is not carried out in this work.

5 Matching Strategy

The last remaining step in our retrieval system is the matching. In the following the basic structure of our matching strategy is developed.

5.1 Low Level Interest Region Matching

The intrinsic goal of interest region matching on descriptor level is to establish reliable region to region correspondences. The nearest neighbor distance ratio (NNDR) provides sophisticated ones although under certain circumstances might be harmful as illustrated in Figure 7. Hereby, due to the local similarity an actually correct matching candidate is discarded. In the case of different trademark instances of the same brand this is especially problematic. These contain partly exactly the same writing and are therefore very likely to block descriptor matches of each other.

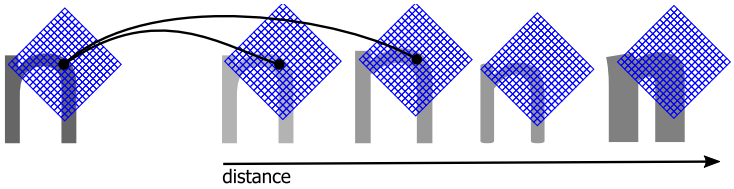


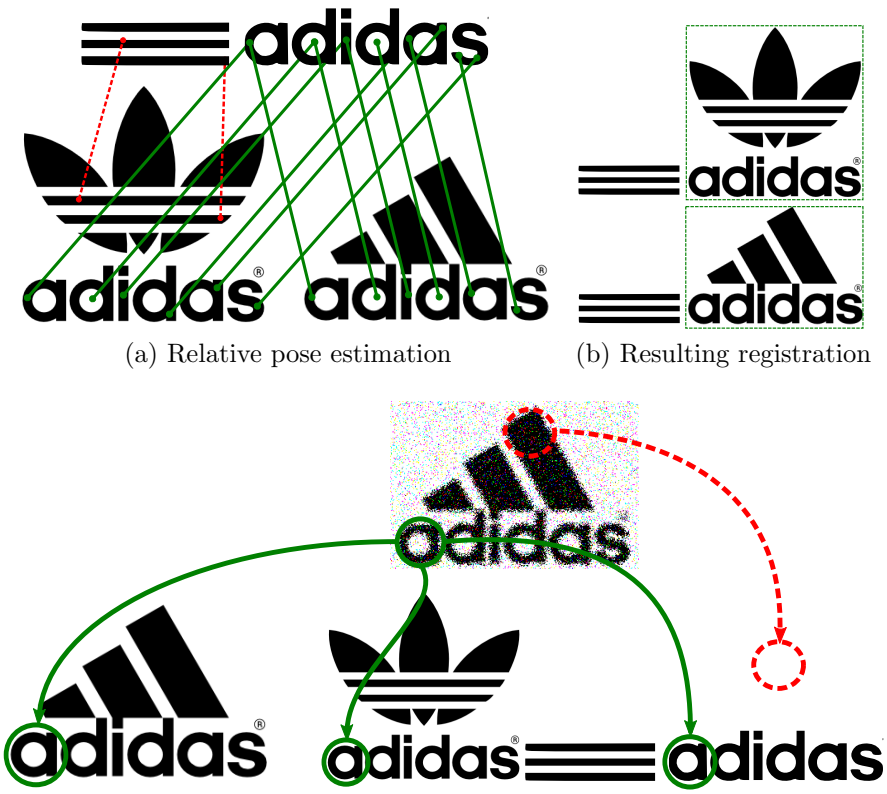
Figure 7: Problems of NNDR. The query descriptor on the left is correctly assigned to its match candidate on the right. As the second nearest neighbor is very close the NNDR discards the correspondence.

Indisputable, the arising problems of NNDR emphasize that also the idea of matching descriptors with direct NN assignment is pursued. In consequence, further investigations should enable us to compare the loss of correct correspondences compared to the increased computational effort due to a more sophisticated verification, caused by a higher outlier rate.

5.2 Creation of Model Families

If using NNDR, different trademark instances of the same brand tend to block descriptor matches of each other. Therefore, we propose to merge the different instances of a trademark which share same graphical or figurative elements into a common representation, termed "model family", to abandon this problem.

To obtain the model families, we register the different trademarks relative to each other by descriptor matching and RANSAC as robust estimator. As model we utilize a similarity transform. Hence, arising incoming matches for a particular training image induce virtual matches, whereby the needed coordinates are obtained by the similarity transform for the other family members. After each family member has pursued the verification (Sec. 5.3) the most likely family member is accepted as matched particular model. Figure 8 illustrates the model family generation steps.



(a) Relative pose estimation (b) Resulting registration

(c) Match redirection

Figure 8: Model family generation steps: (a) Tentative descriptor match correspondences are subject of a robust fit (b). Therefore, the spatial offset, scale, and orientation is known, needed to redirect a match between the different trademarks as shown in (c).

5.3 Intermediate Match Clustering, Pose Estimation and Verification

In the next step, the independent correspondences obtained by the actual descriptor matching are clustered by a Hough transform in pose space which follows [15]. The basic idea is that descriptor matches of a particular class that are consistent with a certain pose hypothesis are identified and subject of a detailed pose estimation. The pose clustering accounts spatial coordinates, scale, and orientation, therefore modeling a similarity transform and serves so as first rough pose estimation.

In our case, the peaks of the Hough transform are not directly subject to an iterative solution for the affine transform parameters. Instead, we utilize Random Sample Consensus (RANSAC) [7] as robust estimator to be able

to deal with more potential false matches. The false matches arise due to NN descriptor matching and coarse Hough bins. The coarse bins should account the detector errors caused by the huge perspective variations and scale changes. For the solution of the affine transform parameters we use a slight modification of the standard RANSAC approach. In addition to the point correspondences we provide further cues such as scale and orientation to determine a robust fit. Basically, similar to the spatial coordinates the scale and orientation of the region can be transferred in both images, where the agreement between query regions and training regions is measured. For the orientation the absolute angle difference is measured; for the scale the ratio is considered. Similar to the distance a specific threshold for each of them accepts or rejects a particular correspondence.

Due to the nature of trademarks and the high outlier rate many incorrect pose hypotheses arise even after Hough transform in the pose space and the first RANSAC estimate. Therefore, a pose hypothesis is subject of further verification steps. With guided-matching additional match correspondences are obtained that support the pose hypothesis and succeeding used to refine the transformation model. According to the refined pose hypothesis the training image is projected and interest regions are extracted and described. Finally, a probability model based on [14] assesses the likeliness for this particular configuration. The model considers, among others, the inlier/outlier probabilities for feature matches and also the feature density in the ROI in the query image and in the projected training image. In contrast to [14], in our approach the interest region inlier/outlier probability is estimated directly based on the guided-matching and the RANSAC estimate. The final probability for the pose hypothesis is thresholded and used to accept or reject a detection. Overlapping probable detections are non-maxima suppressed.

6 Trademark and Logo Retrieval System

Now, having introduced and discussed all modules required to build our system, we now give an overview of the basic structure, which is illustrated in Figure 9. The pipeline is separated into off-line training of the trademarks and retrieval in the test images.

As first step in the training, artificial view sets for the trademarks are synthesized. The needed homographies are determined of manually labeled real world trademark appearances. Succeeding, DoG and Multi-Resolution MSER interest regions are extracted, separately for each view. The SIFT descriptors of the DoG regions are calculated on the texture of the image patch at the identified scale and orientation. In contrast to the MSERs,

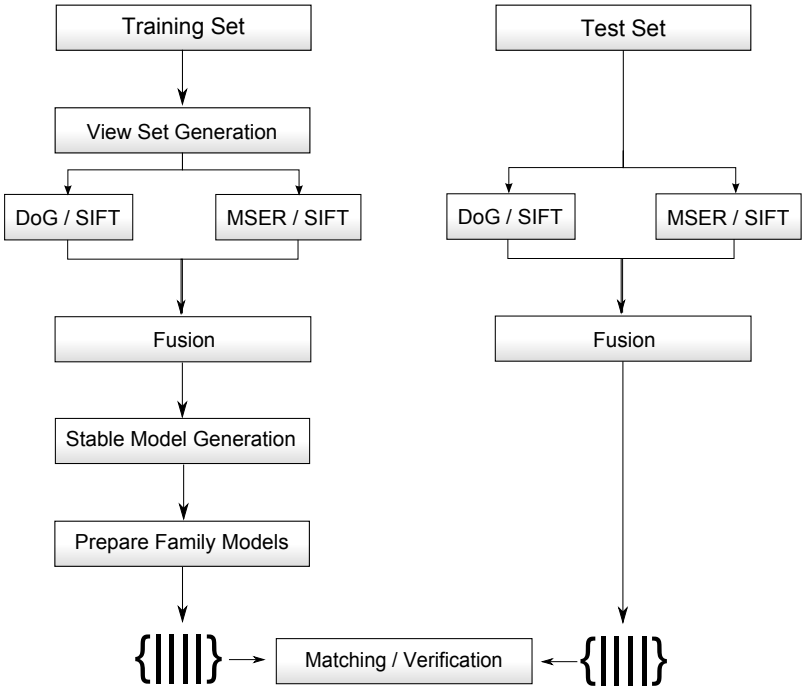


Figure 9: Proposed recognition chain.

hereby the descriptors are calculated on the binarized regions. Therefore, the regions are normalized by LAF covariance construction. Thus, the remaining rotation ambiguity is resolved over the normalized contour distances. To compare the relative size of MSER regions, we use a measure derived of the axis lengths of the respective covariance ellipse. Following, for each trademark a stable model is established. Hereby, we have two choices: First, the most stable descriptors of the view set are iteratively selected, enforcing a minimum distance in feature space. Second, the k-means view tree is constructed. Furthermore, if NNDR matching is employed the trademarks are organized in model families.

Finally, the test images are matched against the trained models. No view set is generated in before and the MSERs are extracted only in a single image resolution. Interest region matching is carried out by matching the test image descriptors on the trademark descriptors. Based on a Hough transform in pose space and RANSAC matches are filtered. Individually for each tentative detection, guided-matching is utilized to further search for interest regions that support the pose hypothesis. The likeliness of the detection is assessed by a probability model. The model considers, among others, the inlier/outlier probabilities for feature matches and also the feature density. Overlapping probable detections are non-maxima suppressed.

7 Experimental Results

In the following, we will show the benefits of the proposed trademark and logo recognition system on a representative test set of 106 still images of high difficulty and variety of the EURO 2008 final game. The task is to detect and to recognize each of the 302 manually annotated trademark occurrences. In total 18 different trademark instances are labeled and organized in 13 families. The dataset contains large perspective changes, partial occlusion, background clutter, motion blur, and in some cases also non rigid transformations. As minimum detection size a minor side length of 15 pixels is enforced. The convenience of a particular feature combination is assessed and discussed compared to the performance of the standard DoG/SIFT approach. As input images for the training trademark database the cropped scene appearances consciously have not been used, since this would trim the approach on this particular data set limiting an objective performance analysis.

7.1 Parameters

Several parameters of the outlined recognition chain influence the respective recall and precision. To allow to compare the applicability of the different feature types and matching configurations, the parameters of the recognition chain are set to their default values. An overview of the used parameters is given in Table 1.

Parameter	**Value**
NNDR threshold	0.8
NN threshold	$\pi/6$
Dist. crit. (Bhattacharyya)	$\pi/6$
Max. # of selected descriptors	200
# of clusters (k-means)	10
Hough location bin size	0.5 of avg. proj. size
Hough orient. bin size	$pi/6$
Hough scale bin subdiv. factor	2
RANSAC orient. thresh.	$pi/6$
RANSAC scale ratio thresh.	0.5 / 2
RANSAC distance thresh.	20 to 5% of proj. size
Probability model thresh.	0.95

Table 1: Recognition chain parameters.

7.2 Model Families Revised

The evaluation of the model family approach on our EURO 2008 test data set provides strong evidence that it is beneficial when matching using NNDR. In total 56 Adidas appearances are labeled in the test set of 106 images. The standard approach delivers only 17 correct detections. In contrast the family model approach delivers clearly a superior performance, with 29 correct detections. Indisputable, this is a strong argument for the use of the model family approach. The effects on selected scene images are illustrated in Figure 10. In the case of the non-family approach it can be seen that the number of correct match correspondences decreases, since the NNDR discards the match candidates which are not distinctive enough for a particular model. With the model family approach the match correspondences are preserved. In consequence, we stick to the model family approach for every trademark database that is matched with NNDR in the succeeding evaluations.

7.3 Comparison of Different Feature Types

The convenience of a particular feature combination is assessed and discussed compared to the performance of the standard DoG / SIFT approach. To further discuss the influence of the matching type the descriptor matching is carried out using both, NN and NNDR matching. To confine the computational costs, due to the vast number of regions present in a typical query image the maximum number of allowed matches for a particular training image is thresholded by 500 for both matching strategies. Table 2 summarizes the results of the experiment. The recall and precision values of the best performing method itemized in the particular trademark families are shown in Figure 11. Illustrative results of the best performing method are shown in Figure **??**.

The MSER regions described by a SIFT descriptor on the region shape turns out to be a reasonable extension to DoG / SIFT. Superior performance is reached by the proposed k-means view tree, which boosts the recall up to about 85% compared to 51% of the baseline method. A further extension by MSERs does not increase the performance. But, if attention is directed on precision our compact stable model approach by descriptor selection, with both DoG and MSER features, is a reasonable choice for a clear performance boost of nearly 22%.

One evident, important finding of our evaluation is that it is feasible for our approach to directly match NN instead of NNDR. In fact, we show that the precision can be maintained largely. It is proven that the method is able

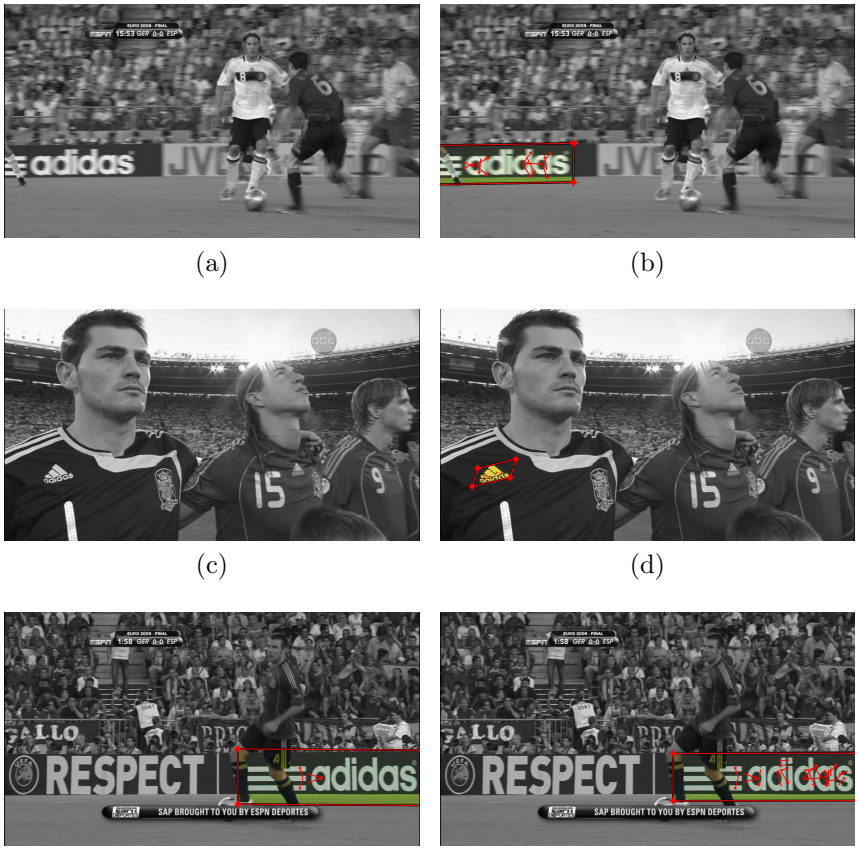


Figure 10: Model family approach (right column) compared to standard approach (left column). The red arrows indicate pose consistent succeeding matches.

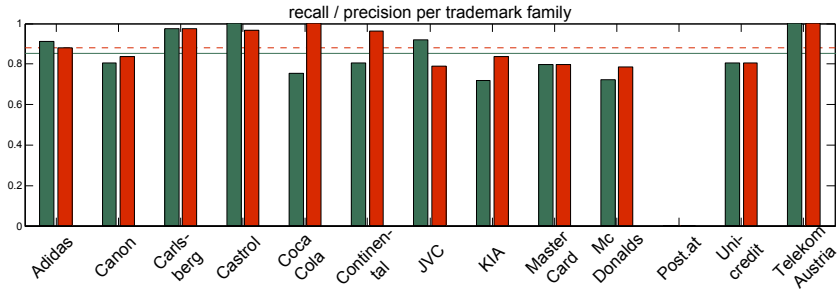


Figure 11: Recall and precision per trademark family. As feature combination a k-means view tree with DoG / SIFT features has been used. The recall is shown in green, precision in red.

to deal with a high amount of false interest region correspondences, due to the exact pose estimation and verification. Therefore, NN matching is possible since recall and precision scale well on increasing trademark database size. Nearly all of the false positives exist in frames with text inserts with similar fonts. Thus, it is assumed that the NN threshold can be furthermore optimized to even enhance the performance, although the exhaustive computational costs remain as bottleneck.

No.	**DoG/SIFT**	**MSER/SIFT**	**SM**	**MT**	**Rec.**	**Prec.**
1	Texture	-	-	NNDR	51,6%	95,6%
2	Texture	-	-	NN	60,5%	93,8%
3	-	Shape	-	NNDR	8,9%	100%
4	-	Shape	-	NN	7,6%	100%
5	Texture	Shape	-	NNDR	61,6%	97,8%
6	Texture	Shape	-	NN	68,4%	93,1%
7	Texture	-	DC	NNDR	61,4%	97,9%
8	Texture	-	DC	NN	65,7%	92,7%
9	Texture	Shape	DC	NNDR	73,3%	97,3%
10	Texture	Shape	DC	NN	78,8%	85,1%
11	Texture	-	tree	NN	84,7%	87,7%
12	Texture	Shape	tree	NN	83,5%	86,4%

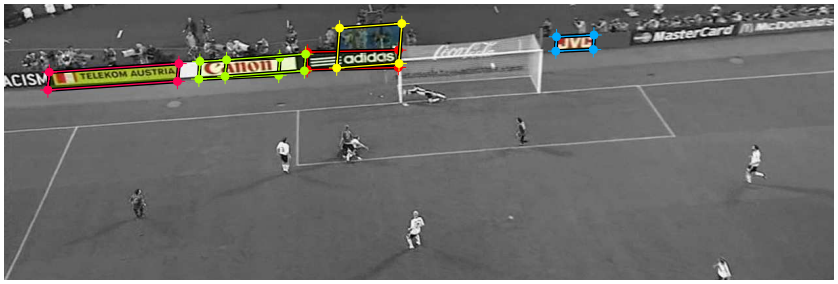
Table 2: Performance comparison of trademark databases with different feature types. Compared to the DoG / SIFT approach the proposed extensions yield a better performance. SM stands for stable model, DC for distance criterion, MT for matching type, R for recall, and P for precision.

8 Conclusion

In this work we presented a specific trademark and logo retrieval system, which allows for automatic efficiency analysis of advertisements. In particular, the goal of this work was to overcome typical problems of existing approaches such as little or low textured trademarks, severe perspective changes, or highly correlated data. These problems are addressed by additionally introducing a complementary feature type (*i.e.*, Multi-Resolution MSER), view-based stable models (by introducing a stable region detector and an efficient data structure), and an efficient matching strategy (*i.e.*, NNDR combined with a clustering of model families). Altogether an efficient (in terms of accuracy, recall, and speed) logo retrieval system is derived, which can be applied in practice. This was illustrated on a competitive real-

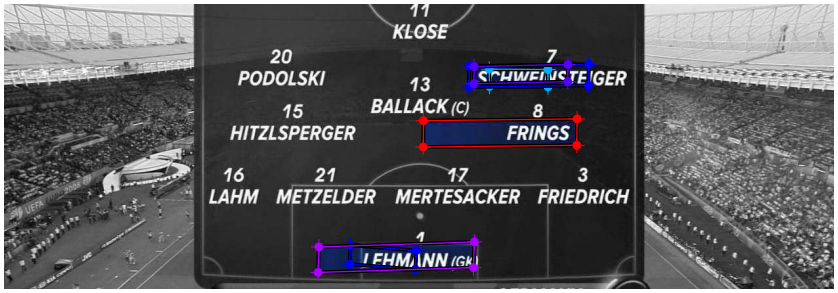
world dataset, *i.e.*, showing frames from the EURO2008 final game, where the results clearly show that compared to a simple SIFT-based detection approach the performance can drastically be improved. A straight forward extension would be to additionally include color information and further complementary region detectors to cope with the high variability in the data.

(a) Occlusions.



(b) Perspective variations.

(c) Perspective variations.



(d) Failure cases, likely in conjunction with text inserts.

Acknowledgement

This work was mainly created in cooperation with Joanneum Reasearch Forschungsgesellschaft mbH. In addition, it was partially supported by the FFG project MDL (818800) under the Austrian Security Research Programme KIRAS.

References

[1] L. Ballan, M. Bertini, A. D. Bimbo, and A. Jain. Automatic trademark detection and recognition in sport videos. In *Proc. IEEE Intern. Conf. on Multimedia & Expo*, 2008.

[2] J. S. Beis and D. G. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.

[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24:509–522, 2002.

[4] O. Chum and J. Matas. Geometric hashing with local affine frames. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 879–884, 2006.

[5] R. den Hollander and A. Hanjalic. Logo recognition in video stills by string matching. In *Proc. IEEE Intern. Conf. on Image Processing*, pages 517–520, 2003.

[6] O. El Badawy and M. Kamel. Shape-based image retrieval applied to trademark images. *Integrated image and graphics technologies*, pages 373–392, 2004.

[7] M. A. Fishler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Tech report 213, AI Center, SRI International*, 1980.

[8] A. Folkers and H. Samet. Content-based image retrieval using fourier descriptors on a logo database. In *Proc. Intern. Conf. on Pattern Recognition*, pages 521–524, 2002.

[9] P.-E. Forssén and D. Lowe. Shape descriptors for maximally stable extremal regions. In *Proc. IEEE Intern. Conf. on Computer Vision*, October 2007.

[10] D. Hall, F. Pelisson, O. Riff, and L. Crowley. Brand identification using gaussian derivative histograms. *Machine Vision and Applications*, 16(1):41–46, 2004.

[11] Y. S. Kim and W. Y. Kim. Content-based trademark retrieval system using a visually salient feature. *Image and Vision Computing*, 16:931–939, 1998.

[12] V. Lepetit, J. Pilet, and P. Fua. Point matching as a classification problem for fast and robust object pose estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages II: 244–250, 2004.

[13] C. Li, G. Cai, and L. Chen. Billboard advertisement detection in sport TV, 2003. Proc. IEEE Intern. Symposium on Signal Processing and its Applications.

[14] D. G. Lowe. Local feature view clustering for 3d object recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 682–688, 2001.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Intern. Journal of Computer Vision*, 60(2):91–110, 2004.

[16] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[17] J. Matas, T. Obdrzalek, and O. Chum. Local affine frames for wide-baseline stereo. In *Proc. Intern. Conf. on Pattern Recognition*, volume 4, pages 363–366, 2002.

[18] G. Medioni, G. Guy, H. Rom, and A. Franois. Real-time billboard substitution in a video stream. In *Proc. Tyrrhenian Intern. Workshop on Digital Communications*, pages 71–84, 1998.

[19] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2161–2168, 2006.

[20] Š. Obdržálek and J. Matas. Object recognition using local affine frames on distinguished regions. In *Proc. British Machine Vision Conf.*, volume 1, pages 113–122, 2002.

[21] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[22] J. Schietse, J. P. Eakins, and R. C. Veltkamp. Practice and challenges in trademark image retrieval. In *Intern. Conf. on Image and Video Retrieval*, pages 518–524, 2007.

[23] A. Watve and S. Sural. Soccer video processing for the detection of advertisement billboards. *Pattern Recognition Letters*, 29(7):994–1006, 2008.