Large-Scale Tattoo Image Retrieval

Daniel Manger
Video Exploitation Systems
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB
Karlsruhe, Germany
daniel.manger@iosb.fraunhofer.de

***Abstract*—In current biometric-based identification systems, tattoos and other body modifications have shown to provide a useful source of information. Besides manual category label assignment, approaches utilizing state-of-the-art content-based image retrieval (CBIR) techniques have become increasingly popular. While local feature-based similarities of tattoo images achieve excellent retrieval accuracy, scalability to large image databases can be addressed with the popular bag-of-word model. In this paper, we show how recent advances in CBIR can be utilized to build up a large-scale tattoo image retrieval system. Compared to other systems, we chose a different approach to circumvent the loss of accuracy caused by the bag-of-word quantization. Its efficiency and effectiveness are shown in experiments with several tattoo databases of up to 330,000 images.**

***Keywords*- content-based image retrieval, biometrics, tattoo images, identification, forensic database**

I. INTRODUCTION

The identification of individuals can be performed using different biometric modalities. Although fingerprints play the most important role in forensic and law enforcement agencies, research in biometrics considers many other modalities such as face, iris, veins or tattoos and other body modifications. For tattoos, the ANSI/NIST-ITL 1-2000 standard defines the eight classes human, animal, plant, flags, objects, abstract, symbols and other [1]. Despite the fact that the standard contains another 80 subclass labels, the matching based on manually assigned class labels is subjective, time-consuming and has other limits as, for instance, tattoos have a large intra-class variety and cannot always be assigned to only one (sub)class. As hardware abilities and image retrieval algorithms rapidly advanced in recent years, appearance–based tattoo matching dealing with images of tattoos gets more and more attention. The advantages of image retrieval methods are obvious: every tattoo can be regarded as a separate class making it possible to distinguish for example different dragon tattoos based on their different visual appearance.

The aim of the system presented in this work being a part of the EU-funded research project FAST and efficient international disaster victim Identification (FASTID) [4] is to support disaster victim identification based on novel tattoo retrieval methods.

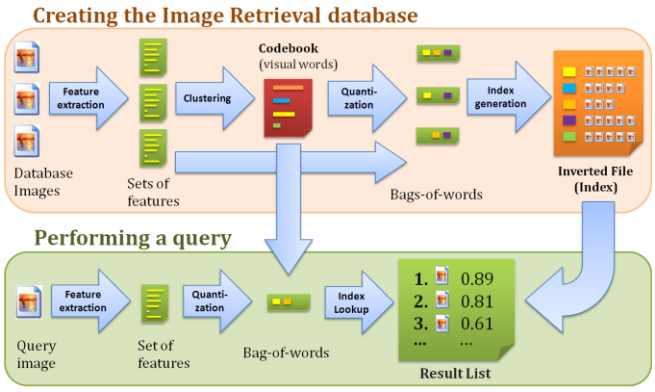

Fig. 1: Basic setup of a content-based image retrieval system which quantizes features into visual words to create an inverted file.

The structure of this paper is as follows: in section II, the typical architecture of image retrieval systems is outlined. Section III presents related work in the context of tattoo retrieval stating our main contributions. Section IV then describes our system setup. Our experiments carried out are presented in section V followed by a brief conclusion.

II. CONTENT-BASED IMAGE RETRIEVAL

A. Comparing image content

The aim of content-based image retrieval systems is to compare images with respect to their content. To this end, local image regions are compared using local features. Local features like the popular Scale-Invariant Feature Transform (SIFT) [16] are used in many different topics of computer vision. They typically detect repeatable salient regions in an image and subsequently encode their local image appearance in a descriptor. Given the two sets of descriptors of two images, similar regions in both images can be searched by determining descriptors which are similar in descriptor space, which is for SIFT usually 128 dimensional. Typically, distances are calculated by L2 norm and a threshold is applied on the distance or on the ratio of closest to second

Dataset	Content[1]	Images	Features	Features per img.	Index size	Rank-1% BOW	Rank-1% best[2]	Rank-1% best[2] RR20[3]	Retrieval time best[4]
8k	T	8,425	14.4 M	1,700	180 MB	38.1%	87.5%	91.1%	176ms
10k	T	9,631	13.8 M	1,400	170 MB	38.4%	84.9%	89.9%	145ms
330k	T+B	327,049	333.7 M	1,000	4 GB	35.7%	78.4%	84.1%	5s
ESP10k	R	9,631	4.8 M	500	68 MB	48.9%	85.4%	90.9%	125ms
2×417	T	834	1.7 M	2,000					
[13]	T+R	100,000	7.8 M	78		(71.1%)	(77.3%)		(270ms)

Table 1: Characteristics of datasets and results. [1] T=Tattoos, B=Body modifications (e.g. scars, marks), R=Random content.
[2]Best denotes the technique used to circumvent the loss of performance due to BOW quantization, i.e. HE and WGC in this work and ensemble ranking in [13]. [3]With applying a re-ranking of the top 20 images using the original (not quantized) features. [4]For an image containing 2000 features without re-ranking, time for feature extraction and feature quantization not included. Please note that this work and [13] use different query and database images. Consequently, the retrieval accuracy and time (put in brackets) cannot be directly compared to this work.

closest distance. The similarity of two images is then often calculated as the number of matching features.

B. *Quantization of features*

For matching sets of descriptors, various heuristic algorithms have been proposed which can lead to an impressive speedup while sacrificing not too much of the descriptors discriminance [19],[21]. Nevertheless, in large-scale CBIR systems with thousands or millions of images, a pair-wise image comparison of the query image with every image of the database becomes infeasible. Besides, the memory consumption of the image features and their processing during one query prohibit a direct matching of descriptors sets. To solve this, the bag-of-words (BOW) representation has been proposed [26], which quantizes the features by assigning every feature to one element of a set of feature representatives called visual words. Thus, the image matching can be performed with text retrieval methods analyzing the common visual words of images. The set of visual words termed codebook or visual vocabulary is commonly obtained by clustering an independent set of features. Using large codebooks, the representation of an image becomes a very sparse vector indicating the occurring visual words. This sparsity can be exploited by inverted files which store for every visual word a list of references to the images containing at least one feature corresponding to that visual word. Figure 1 summarizes the basic components of an image retrieval system.

C. *Similarity score*

As rare visual words are assumed to be more discriminative, the similarity of two images given the two BOW vectors is commonly calculated using the *tf-idf* scheme [26]. It weights the BOW vectors according to both the local frequency (within the image) and the global frequency (within the entire database). In all experiments in this paper, we use the similarity function of [25] which is the cosine angle between the weighted BOW vectors which equals the L2 normalized dot product of the vectors. See [10] for details.

III. Related Work

Early approaches for using CBIR methods for tattoo retrieval have focused on low-level features like color, shape and texture [9] or Fourier shape descriptors [7]. Being extracted on the whole image, their main shortcoming is that they often need preprocessing steps to extract the relevant foreground region of the tattoo. Moreover, their discriminability in tattoo retrieval is limited which leads [12] to apply a rank-based distance metric learning. [11] introduces local features for tattoo retrieval and demonstrates their superiority to low-level features. In [8], the incorporation of label information (tattoo type and body location) is shown to improve the performance. However, all these approaches perform a direct matching of features i.e. a linear scan of the database is required which prohibits them from being used in large scale systems. As an answer to that problem, the bag-of-words model [26] is proposed in [15] in combination with a feature quantization method focusing on the computational cost of feature clustering.

The loss of performance due to the BOW quantization in tattoo retrieval has been recently addressed in [13]. They propose an ensemble of models. More precisely, ten different BOW models are generated using different initializations of the K-means clustering in the codebook generation step. Afterwards, an unsupervised learning algorithm is presented which learns weights for combining the models into one system fusing the retrieved ranks of the ten subsystems. Although the rank-1 accuracy showed to increase by 6%, there is a computational overhead in using multiple BOW models.

The work most similar to this paper is [13]. However, we use a different approach to circumvent the loss of accuracy caused by the bag-of-word quantization namely Hamming Embedding (HE) [10] and Weak Geometry Consistency (WGC) [10]. Both techniques have shown a significant improvement of performance in large scale image retrieval. While so far mainly tested in standard datasets containing images of buildings or scenes, we show in this work that

they also can greatly enhance the performance in tattoo retrieval. To evaluate the benefit, we use databases containing up to 330,000 images of tattoos and other body modifications. To our knowledge, our system processing over 300 million features demonstrates tattoo retrieval using the largest content relevant database as distracters.

IV. Large Scale Image Retrieval

We build on the basic setup of a CBIR system described in the second section and displayed in Fig.1. However, in contrast to [13], to counter the loss of performance due to quantization, we don't use ensemble techniques, but use Hamming Embedding (HE) [10] and Geometry Consistency (WGC) [10].

A. Hamming Embedding

Hamming Embedding extends the information of a quantized feature x by a d_b-dimensional binary signature $b(x) = (b_1(x), \dots, b_{d_b}(x))$. The idea behind this is, that the hamming distance $h\big(b(x), b(y)\big) = \sum_{i=1}^{d_b} |b_i(x) - b_i(y)|$ between the signatures of two features which are quantized to the same visual word i.e. lying within the same Voronoi cell approximates the Euclidean distance of the features. To obtain the binary signature of a feature assigned to visual word, its descriptor in Euclidean space is first projected to the d_b-dimensional space using an orthogonal projection matrix. Refer to [25] for details on the creation of an appropriate projection matrix. The b elements of the resulting vector are then compared with respect to the 'typical' distribution of elements within each single dimension. This distribution is solely represented by a set of b median values $m_1 \dots m_b$ for every visual word which are determined by offline projecting a sufficiently large set of independent features. The binary signature finally binary encodes for each dimension $i = 1 \dots b$ either 1 or 0 depending on the element being larger or smaller than the median m_i of that dimension of that visual word.

The similarity scoring can make use of the HE information by introducing a hamming distance threshold for filtering the matches voting for the images in the database. As a consequence, database images can no longer be represented by a BOW vector accumulating the features for every visual word but by a list of quantized features including the HE signature. However, as commonly large codebooks are used, most quantized features only occur once within an image limiting the extra amount of space – apart from the bits needed for HE. We use $d_b = 32$ bit hamming codes and apply a threshold of 18 bits. The matches which pass the HE filtering are additionally weighted Gaussian according to their hamming distance, see [24] for details.

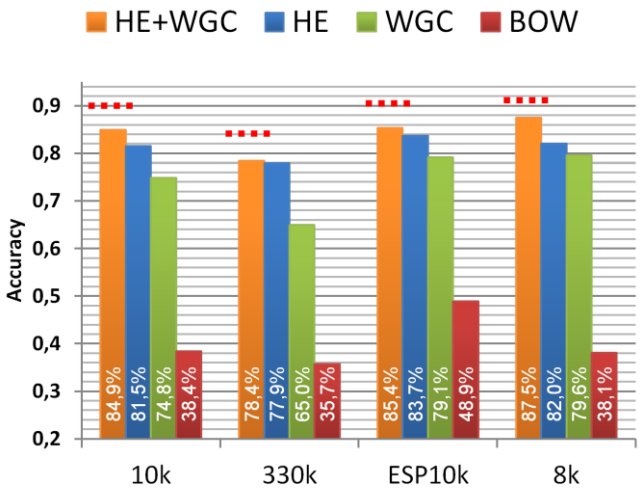

Fig. 2: Rank-1 accuracy using 417 pairs of test images and four different datasets as distracters (without re-ranking). The 8k dataset was used to extract the Codebook. The dotted red lines indicate the results after re-ranking the first 20 images.

B. Weak Geometry Consistency

Up to now, information about the geometric distribution of the features or matches is not used. Many systems make use of this information through estimating a 2D affine transformation based on the matches of two images. Due to complexity, this can only be applied to a subset of the images leading to a re-ranking of the first few images. In contrast to that, Weak Geometry Consistency (WGC) [10] uses geometry information already in the first stage of the retrieval system which implies - as for HE - an extension of the inverted file. The basic idea of WGC is to additionally use the orientation and scale information of matching features. For a pair of matching images, the histogram of the orientation differences of all feature pairs of the matches should have a maximum bin which corresponds to the global rotation of the matching object. The same holds for the scale parameters of the features. However, the scale information is often less reliable and therefore we only use the orientation information in our experiments. To this end, the accumulator which collects the votes for every possible target image is changed to contain 36 bins for orientation differences for every target image. The final score for every image is then given by the maximum value of the bins. To reduce quantization effects, we use the sum of the maximum bin and its two neighboring bins instead.

Using HW and WGC, each feature of the database images results in an entry in the inverted file which contains not only the image number, but also the 32 bit HE code and the orientation information which in our setup sums up to 12 bytes per feature.

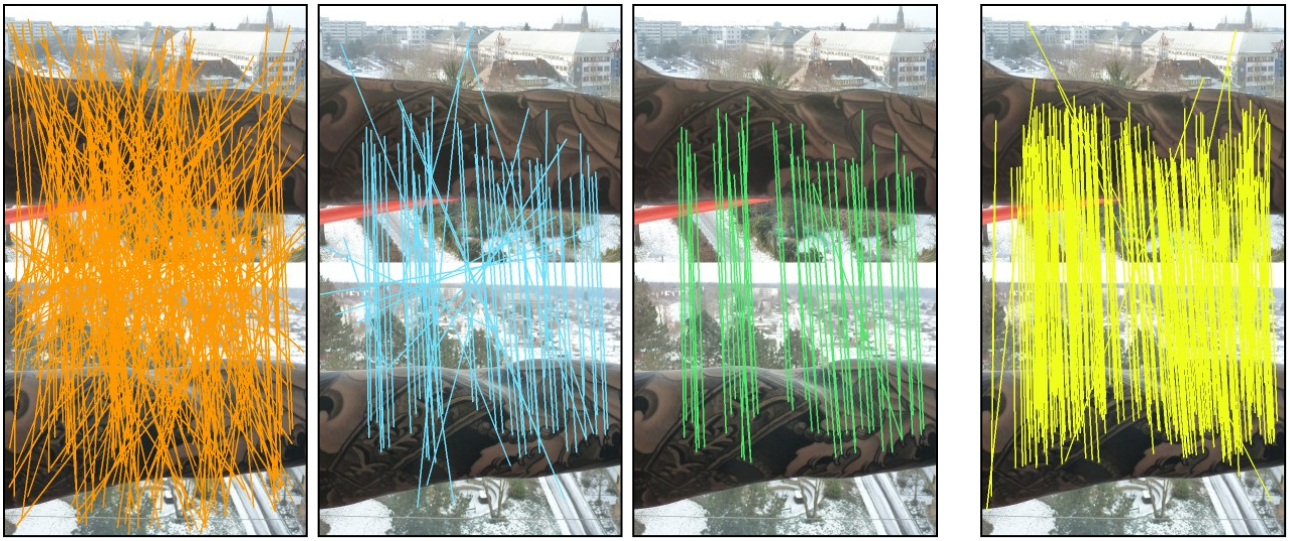

Fig. 3: Matching features of two images showing the same tattoo. From left to right: bag-of-word (BOW) [26] matches, BOW matches after Hamming Embedding (HE) [10] filtering, BOW matches after HE filtering using Weak Geometry Consistency (WGC) [10] and direct matches (using the raw features i.e. without quantization). Originally, 2944 features have been extracted in the upper image and 5422 features in the lower image. As can be seen, HE and WGC clearly succeed in eliminating all false matches on the background caused by the quantization thus enabling the system to work with images in which the tattoos' Regions Of Interest are not available.

V. EXPERIMENTS

A. Datasets

Unlike for face recognition, there are no common evaluation databases to assess tattoo retrieval systems. We therefore downloaded tattoo images from different tattoo websites:

- *8k-DB*: 8,425 images from tattoodesign.com. These images have been used to generate all vocabularies and the HE medians used in this work.
- *10k-DB*: 9,631 images from eviltattoo.com. These images are used as database images.
- 2 images each of 417 tattoos (i.e. 834 images in total) from wildcat.de. These two different images for each of the 417 tattoos allow a performance evaluation using one image of every tattoo as query image and the other as database image to be retrieved.
- *330k-DB:* we have access to 327,049 images of tattoos and other body modifications collected by the German police which is used for large scale tests.
- *ESP10k-DB*: similarly to [13] we used 9,631 images from the ESP game [3] available from [5] (the 9,631 images yielding the most features) to investigate the importance of the context of the images used as distractors.

Please note that the 417 images used as queries refer to 417 corresponding images which are in contrast to [9],[11],[12] not synthetically generated but show a variety of challenging real-world transformations: (1) large scale and certain viewpoint differences, (2) a lot of background clutter, (3) different stadiums of tattooing process, etc. More details of the datasets are given in Table 1. Due to privacy and copyright issues, images of the databases above cannot be shown to demonstrate the algorithms. Instead, we present own images and images of the Centre for Anatomy and Human Identification (CAHID) at the University of Dundee, which is currently establishing an image database of body modifications [2]. Images are chosen to illustrate the same situation.

B. Evaluation setup

By querying all of the 417 test images, the ranks of the respective true corresponding images in the result list are gathered yielding a histogram which represents the number of images for all occurring ranks. Subsequently, the histogram is accumulated leading to the well known Cumulative Match Curve (CMC) [18] commonly used in image retrieval evaluation. The curve specifies for every rank the percentage of the correct corresponding images which have been presented by the system up to that rank.

We extract SIFT features [16] for all images and use hierarchical K-Means clustering [21] for generating a visual vocabulary of size $7^6 = 117649$ with the images from 8k-DB. As a baseline, we use the BOW model [26] and the similarity scoring described in Section II. We use multiple assignment [23] and assign each feature of the query image to the closest two visual words.

C. Results

With 10k-DB as distractors, the baseline system retrieves 160 of the 417 test images on the first rank (43%). Expectably, the performance decreases for the larger 330k-DB (see Fig.2). However, both Hamming Embedding and Weak Geometry Consistency can partly compensate for the loss off accuracy caused by the quantization. In combination, they are able to push the rank-1 accuracy from 38% to 85% in the 10k-DB and from 36% to 78% in the 330k-DB. Fig.3 illustrates the benefits of applying HE and WGC to a matching image pair in terms of its filtering capabilities. All incorrect BOW matches occurring from background clutter are filtered and only one incorrect match on the tattoo is left. Thus, HE and WGC enable the system to be used for images without providing any annotation or segmentation of the tattoo location. For sake of comparison, we include the direct feature matches calculated by the Euclidean distances of the descriptors and using a threshold of 0.5 for the ratio of closest to second closest match [16]. Note, that it yields five times more matches but needs more than ten times the memory compared to the quantized version with HE code and orientation information.

Comparing the baseline performance of the 10k and ESP10k dataset clearly shows that using images from a different domain can limit the meaningfulness of large scale tests. Even though we found a few tattoos which seem to be part of both databases and therefore possibly affect the rank of the corresponding truth test images, the tattoo images in 10k-DB are more distracting than the images of the ESP10k dataset which makes the matching job easier for the ESP10k case.

D. Re-ranking

Fig. 5 shows the Cumulative Match Curve of the 10k-DB and 330k-DB experiments using HE and WGC (dotted lines). The significant increase of the curve within the first 20 images indicates that they tend to be quite similar. To further improve the performance of our system, we thus applied a subsequent re-ranking step which performs a matching based on the original features. The images are re-ranked according to the number of matches with the query image (for example 257 in the rightmost image of Fig.3). This improves the rank-1 performance for the 10k-DB by 5.0% and for the 330k-DB by 5.7% (see solid lines). Table 1 summarizes the results obtained with the different datasets.

E. Runtime

The calculation of the hamming distance of two signatures (corresponding to a binary XOR operation followed by counting the resulting nonzero bits) can hugely be speeded up using SSE 4.2 processor extensions [6]. Moreover, the HE filtering leads to a smaller number of matches contributing to the score of the images. Both circumstances make our implementation with HE slightly faster than without HE (see Fig.4). Performing a query in the 10k-DB with an image having 2,000 features takes 145ms (165ms respectively). For the 330k-DB it takes about 5s (6s respectively). The values are measured without feature extraction and without feature quantization. All experiments have been performed on an Intel i7-930 using 4 cores with 2.8 GHz and 8 GB of main memory.

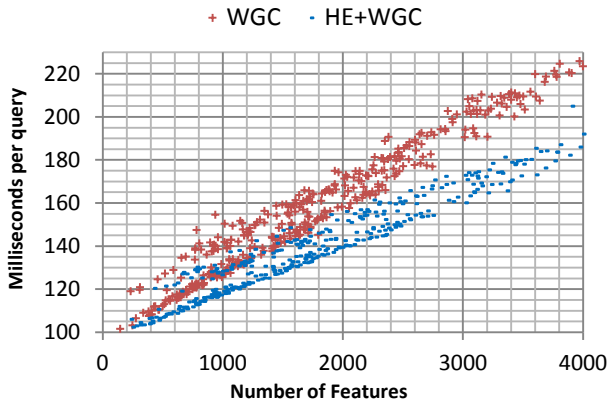

Fig.4: Retrieval times for the 417 query images using the 10k-DB (without feature extraction and without quantization). Due to its filtering capabilities, the integration of HE leads to a slightly reduced retrieval time compared to WGC only although it involves computational overhead.

VI. Conclusion and Future Work

We presented a tattoo retrieval system which builds upon recent image retrieval techniques to ensure scalability towards large databases containing hundreds of thousands of images. Given a query image, the system retrieves corresponding images within a matter of seconds searching in a corpus containing more than 300,000 tattoo images.

The images which could not be retrieved by the system within the first 20 ranks mainly show a large difference with respect to the point of view which leads to serious affine transformations. Especially, when large tattoos on arms or legs are photographed, the two different views often show only a very small overlap in which features can be matched. See Fig.6 for an example image. We therefore plan to further optimize the system by using features capable of dealing with affine transformations [17].

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 242339 (FASTID project [4]). Some images used in this article are kindly provided by the Centre for Anatomy and Human Identification (CAHID) at the University of Dundee, as part of the FASTID body modification research [2].

REFERENCES

[1] ANSI/NIST-ITL 1-2007 standard: Data Format for the Interchange of Fingerprint, Facial, & Other Biometric Information (2007).

[2] Body Modification Research. Centre for Anatomy and Human Identification (CAHID) at the University of Dundee. http://www.bodymodresearch.com/

[3] ESP Game. http://www.gwap.com/gwap/gamesPreview/espgame/

[4] FAST and efficient international disaster victim IDentification (FASTID) Project Website. http://www.interpol.int/Projects/FASTID

[5] John Langford Website. http://www.hunch.net/~jl/

[6] Intel® SSE4 Programming Reference.2007 Intel Corporation.

[7] Acton, S. T., & Rossi, A. (2008). Matching and retrieval of tattoo images: active contour CBIR and glocal image features. In Image Analysis and Interpretation, 2008. SSIAI 2008. IEEE Southwest Symposium on (pp. 21-24).

[8] Jain, A. K., Lee, J. E., Jin, R., & Gregg, N. (2009). Content-based image retrieval: An application to tattoo images. In Image Processing (ICIP), 2009 16th IEEE International Conference on (pp. 2745-2748).

[9] Jain, A., Lee, J. E., & Jin, R. (2007). Tattoo-ID: automatic tattoo image retrieval for suspect and victim identification. Advances in Multimedia Information Processing--PCM 2007, 256-265. Springer.

[10] Jegou, H., Douze, M., & Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In European Conference on Computer Vision. ECCV 2008 (pp. 304-317)

[11] Lee, J. E., Jain, A. K., & Jin, R. (2008). Scars, marks and tattoos (SMT): Soft biometric for suspect and victim identification. In Biometrics Symposium, 2008. BSYM'08 (pp. 1-8).

[12] Lee, J. E., Jin, R., & Jain, A. K. (2008). Rank-based distance metric learning: An application to image retrieval. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (pp. 1-8).

[13] Lee, J. E., Jin, R., & Jain, A. K. (2010). Unsupervised Ensemble Ranking: Application to Large-Scale Image Retrieval. In Pattern Recognition (ICPR), 2010 20th International Conference on (pp. 3902-3906).

[14] Lee, J. E., Jin, R., Jain, A. K., & Tong, W. (2011). Image Retrieval in Forensics: Tattoo Image Database Application. IEEE MultiMedia, 40-49. Published by the IEEE Computer Society.

[15] Li, F., Tong, W., Jin, R., Jain, A. K., & Lee, J. E. (2009). An efficient key point quantization algorithm for large scale image retrieval. In Proceedings of the First ACM workshop on Large-scale multimedia retrieval and mining (pp. 89-96).

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

[17] Mikolajczyk, K., & Schmid, C. (2004). Scale & affine invariant interest point detectors. IJCV, 1(60), 63-86.

[18] Moon, H. (2001). Computational and performance aspects of PCA-based face-recognition algorithms. PERCEPTION-LONDON-, 30(3), 303-322. PION LTD.

[19] Muja, M., & Lowe, D. G. (2009). Fast approximate nearest neighbors with automatic algorithm configuration. In International Conference on Computer Vision Theory and Applications (VISSAPP'09).

[20] Nister, D., & Stewenius, H. (2006). Scalable Recognition with a Vocabulary Tree. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06), 2161-2168. doi: 10.1109/CVPR.2006.264.

[21] Nister, D., & Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on (Vol. 2, pp. 2161-2168).

[22] Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on (pp. 1-8).

[23] Philbin, J., Isard, M., Sivic, J., & Zisserman, A. (2008). Lost in Quantization : Improving Particular Object Retrieval in Large Scale Image Databases. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008.

[24] Schmid, C. (2009). On the burstiness of visual elements. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009., 1169-1176.

[25] Schmid, C. (2011). Improving bag-of-features for large scale image search. In International Journal of Computer Vision (2010), 316-336.

[26] Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on (pp. 1470-1477).

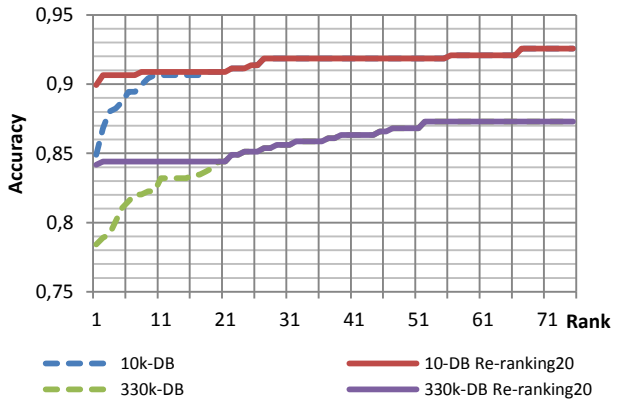

Fig.5: Re-ranking the first 20 images retrieved by the system can increase the performance in both the 10k and 330k dataset.

Fig.6: Limits of using non affine invariant features: Large changes in viewpoint on arms or legs can lead to a small overlap region which in this case after HE filtering only contains one correct match.