EVALUATION OF CLUSTERING CONFIGURATIONS FOR OBJECT DETECTION USING SIFT FEATURES

Fernández Robles, L.; Castejón Limas, M.; Alfonso Cendón, J.; Alegre Gutiérrez, E.

Universidad de León

Scale-Invariant Feature Transform (SIFT) features have been widely accepted as an effective local keypoint descriptor for its robust description of digital image content. This method extracts distinctive invariant features from images that can be used to perform reliable matching between different views of an object. Object recognition proceeds by matching individual features to a database of features from known objects using a fast nearest-neighbour algorithm, followed by a Hough transform to identify clusters belonging to a single object, and finally performing verification through least-squares solution for consistent pose parameters. Nonetheless, reasoning for the choice of this clustering approach is not provided and a lack of its theoretical insight is noticed. Here, we present and evaluate different configurations for clustering sets of keypoints according to their pose parameters: x and y coordinates location, scale and orientation based on Lowe's approach.

Keywords: *ASASEC; SIFT; Clustering*

EVALUACIÓN DE CONFIGURACIONES DE CLUSTERING PARA LA DETECCIÓN DE OBJETOS UTILIZANDO CARACTERÍSTICAS SIFT

Las características SIFT del inglés, Scale-Invariant Feature Transform, se consideran eficientes descriptores locales de puntos clave debido a su robusta descripción de imágenes digitales. Para realizar el reconocimiento de objetos primero se determinan las correspondencias de las características individualmente con una base de características de objetos conocidos utilizando un algoritmo de vecinos cercanos, a continuación se realiza una transformada Hough para identificar clusters pertenecientes a un solo objeto y, finalmente, se lleva a cabo una verificación de la consistencia de la pose del objeto a través de mínimos cuadrados. Sin embargo, no se provee un razonamiento teórico o práctico sobre la elección de este enfoque. Aquí, presentamos y evaluamos diferentes configuraciones para realizar el clustering de conjuntos de puntos clave de acuerdo a los parámetros de pose: coordenadas de localización x e y, escala y orientación basados en el enfoque de Lowe.

Palabras clave: *ASASEC; SIFT; Clustering*

Correspondencia: l.fernandez@unileon.es

1. Introduction

Object retrieval aims at retrieving images that contain objects similar to the query object captured in the region of interest (ROI) of a query image. When the object retrieval system is based on query by example, the user choses an image of interest and then selects a bounding box in that image, which conforms the ROI, holding the query object or object of interest. Straightaway, the ROI has to be described and this feature representation is used to match images or videos in an image or video database. Changes in pose, scale, orientation, illumination, rigidity, cluttered background or occlusion, among others, make object retrieval a challenging task (see Figure 1).

Figure 1: Example of images containing the same object, a blue toy car.

Note: Changes in pose, scale, orientation, illumination and cluttered background can be noticed making the object retrieval task very challenging.

Recent object description approaches rely on local features rather than in global descriptors since local description can reliably detect highly distinctive keypoints of an image. Therefore, a single feature can be correctly matched with high probability against a large database of features, providing a basis for object and scene recognition. Another advantage is the capability for still retrieving images with occluded query objects. The development of image matching by using a set of local interest points was definitively efficient when Lowe (1999) presented SIFT, introducing invariance to the local feature approach. Later, Bay, Tuytelaars and Van Gool (2006) outperformed previously proposed schemes with respect to repeatability, distinctiveness, robustness and speed. Recently, Özuysaletal et al. (2010) showed a fast key point recognition method using random Ferns which avoids the computationally expensive patch preprocessing by using hundreds of simple binary features. Following this idea and due to the need of running vision algorithms on mobile devices with low computing power and memory capacity new approaches are been developed. The Binary Robust Independent Elementary Feature (BRIEF) (Calonder et al., 2010), the Oriented Fast and Rotated BRIEF (ORB) (Rublee et al., 2011), the Binary Robust Invariant Scalable Keypoints (BRISK) (Leutenegger, Chli, and Siegwart, 2011) and the Fast Retina Keypoint (FREAK) (Alahi, Ortiz and Vandergheynst, 2012) are good examples. Their stimulating contribution is that a binary string obtained by simply comparing pairs of image intensities can efficiently describe a keypoint, for example an image patch.

In order to rank all the images of the dataset one possibility would be to take into account the distance of the closer match between the ROI and every image in the dataset. However this could lead to two kinds of errors. First, the local surroundings of two keypoints could be very similar even when they belong to different objects. Secondly, unfortunately a bounding box selection makes that together with the query object there could be other (partial) objects or cluttered background in the ROI. These non-interest regions can produce non-relevant keypoints that could lead to closer matches. Lowe (2004) suggests to consider at least 3 features correctly matched from each object for reliable object retrieval using a Hough

transform to identify clusters belonging to a single object, and finally performing verification through least-squares solution for consistent pose parameters.

ASASEC (Advisory System Against Sexual Exploitation of Children) is a European research project whose goal is to provide a technological solution to help the fight against child pornography. One of the most challenging tasks in this kind of environments consists of retrieving objects from huge amounts of image and video datasets that are proven to come from images or videos containing scenes of children exploitation. Relating different scenes could help to understand and demonstrate complex legal cases. Object matching interpretation becomes then a crucial task. In this paper we compare the performance when retrieving images using just the distance of the closest pair of keypoints and when using Lowe's (2004) proposal using the vote of at least three points by means of the Hough transform and least-squares verification.

The rest of the paper is organized as follows: After reviewing object recognition using Invariant Local Features (ILF) and the main ideas related with model fitting in section 2, we introduce our experiments and discussion in section 3. Section 4 brings the most important conclusions and recommendations for future work.

2. Object recognition using Invariant Local Features

2.1 Invariant Local Features (ILF) and object recognition

As it has been pointed out in the introduction, since the Lowe's paper (Lowe, 1999) the computer vision community has been very active presenting improvement after improvement based on SIFT method. Although over a decade old, Lowe's algorithm has proven very successful in a number of applications using visual features and mainly object recognition. The main problem associated with it has been the large computational burden imposed by its high dimensional feature vector what, in recent years, led to the emergence of new proposals mainly focused on obtaining equally robust descriptors but more computationally efficient. The first of a series was FAST (Rosten, 2006) who uses a machine learning approach to derive a feature detector similar a Harris, SUSAN or DoG but much more faster that them, but with the disadvantage that is not very robust to the presence of noise.

A step farther was proposed by Calonder et al. (Calonder, 2010). They use binary strings as a feature point descriptor, called BRIEF that is highly discriminative even when using few bits and can be computed using simple intensity difference test. Another big advantage of BRIEF is that the matching can be performed by using the Hamming distance, which is more efficient to compute that the Euclidean distance employed in most of the invariant local features detectors. Both aspects convert BRIEF in a faster descriptor in construction and matching but as it is not invariant to large in-plane rotations it is not suitable for object recognition task that are similar to the problem we are facing in the ASASEC project.

Another faster than the classical SIFT and SURF (Bay, 2006) but at comparable matching performance is the BRISK detector (Leutenegger, 2011). BRISK relies on a configurable circular sampling pattern from which it computes brightness comparison to form a binary descriptor string. Its detection methodology is inspired in the adaptive corner detector proposed by Mair (Mair, 2010) for detecting regions of interest in the image. Their AGAST is essentially an extension for FAST (Rosten, 2006), proven to be a very efficient basis for feature extraction. With the aim of achieving invariance to scale BRISK goes a step further by searching for maxima not only in the image plane, but also in scale-space using the FAST score s as a measure for saliency.

An evolution of the above-mentioned methods is the ORB descriptor (Rublee, 2011) that builds its proposed feature on FAST and BRIEF, standing its name for Oriented FAST and Rotated BRIEF (ORB). Their authors address several limitations of these techniques, mainly the lack of rotational invariance present in BRIEF. They add a fast and accurate orientation component to FAST and, at the same time, they present an efficient way to compute the oriented BRIEF features. Furthermore, the ORB descriptor use a learning method for de-correlate BRIEF features under rotational invariance, leading to better performance in nearest-neighbour applications. ORB was evaluated using two datasets: image with synthetic in-plane rotation and added Gaussian noise, and a real-world dataset of textured planar images captured from different viewpoints. As their authors pointed out, ORB outperforms SIFT and SURF on the real-world dataset, both the outdoor and the indoor one what make this method a good choice for object recognition.

As this work is based in the SIFT descriptor, next section elaborates on describing the main steps that are necessary to obtain a SIFT keypoint vector.

2.2 Scale Invariant Feature Transform (SIFT)

In his seminal paper, Lowe (Lowe, 1999) presented a method for image feature generation called the Scale Invariant Feature Transform (SIFT). The advantage of this approach it that it transforms an image into a large collection of local feature vectors, each of which is invariant to image translation, scaling, and rotation.

The scale-invariant features are efficiently identified by using a staged filtering approach.

i) *Identification of key locations in scale space.*

The first stage identifies key locations in scale space that are invariant to image translation, scaling and rotation, and also robust to noise and small distortions. To achieve rotation invariance SIFT selects key locations at maxima and minima of a difference of Gaussian function applied in scale space that is computed by building an image pyramid with resampling between each level.

ii) *Stability of the key features.*

The stability of the previous keypoints is obtained through different operations. Image gradients and orientations are used to characterize the image at each key location. The gradient is computed as the pixel differences and robustness to illumination change is enhanced by thresholding the gradient magnitudes at a value of 0.1 times the maximum possible gradient value.

To ensure invariant to rotation, to each key location a canonical orientation is assigned. The orientation for each key point is determined by the peak in a histogram of local image gradient orientations. The orientation histogram is created using a Gaussian-weighted window whose weights are multiplied by the thresholded gradient values and accumulated in the histogram at locations corresponding to the orientation.

iii) *Description of the local regions surrounding the key points*

Given a stable location, scale and orientation for each key, SIFT describes the local image region in a manner invariant to these transformations. This method represents the local image region with multiple images representing each of a number of orientations (orientation planes). Each orientation plane contains only the gradients corresponding to that orientation,

with linear interpolation used for intermediate orientations. Each orientation plane is blurred and resampled to allow for larger shifts in positions of the gradients.

SIFT implements this approach in a very efficient way by using the same precomputed gradients and orientations for each level of the pyramid that were used for orientation selection. For each keypoint, they use the pixel sampling from the pyramid level at which the key was detected. The pixels that fall in a circle of radius 8 pixels around the key location are inserted into the orientation planes. The orientation is measured relative to that of the key by subtracting the key's orientation. In his experiments, Lowe uses 8 orientation planes, each sampled over a 4 x 4 grid of locations, with a sample spacing 4 times that of the pixel spacing used for gradient detection.

In order to sample the image at a larger scale, the same process is repeated for a second level of the pyramid one octave higher. This time, SIFT uses a 2x2 sample region, therefore almost the same image region is examined at both scales avoiding this way undesirable effects coming from occlusions. The final SIFT feature vector comes from a 4x4 array of histograms with 8 orientation bins in each, having a length of 4x4x8 = 128 element feature vector for each keypoint.

2.3 Model fitting

The following step after obtaining the data around the key points and their local features might be to extract useful information by using data-mining techniques (Han et al. 2006). This approach seems to be fitted for this scenery due to the large amounts of information that quite soon a medium size collection of pictures might provide. Moreover, there is hidden information within the bulk of key points and their descriptors that might be extracted in order to achieve a better knowledge of the nature of the objects captured in the analysis. That seems to be mandatory for decision-making processes based on an evidence criteria, as recommended by ISO standards.

Amongst the different procedures that these data may undergo, we consistently find of the greatest usefulness the following stages:

- Exploratory data analysis
- Outlier identification
- Variable selection
- Robust model building
- Model testing with new data from fresh images

The former might summarize a general data-mining scheme for computer vision and pattern recognition, at least to the degree where reliable higher information might be obtained.

The first step, *exploratory data analysis*, provides a visual comprehension of the difficulties that the numerical algorithms might have to tackle in order to discern the number of different behaviors ---in computer vision, different objects---, occurring in the pictures data set. A well-trained practitioner might find suggestive elements that might support later results as provided by the numerical algorithms or, on the contrary, help them search further with different techniques should the results be inconclusive, proving that particular technique not accurate enough for the purpose intended.

The second step, *outlier identification*, highlights those samples that do not follow the pattern of the majority. Traditionally, these outliers have been nominated for removal in order for simpler algorithms manage to obtain the model representing the majority of the cases

recorded. Nevertheless, it is sensible to investigate the origin of these outliers, as many times that will provide knowledge about particular events of special relevance. A more coherent approach in model fitting integrates those outliers in the model definition ensuring that the model building process is robust enough to manage this information.

The third step, *variable selection*, plays a central role in computer vision for pattern recognition. Given the various and sundry different descriptors proposed by researchers in computer vision during the last decades, the practitioner must select those that will perform the particular task at hand more efficiently. Variable selection can be performed by traditional ways: adding and removing the input variables and checking whether that improves the results according to some indicator; or following more powerful approaches where evolutionary computation, e.g. genetic algorithms, are especially useful.

The fourth step, *robust model building*, is responsible for providing a prediction model that works well not only with the training data but also with new data from fresh images not considered on the determination of the parameters of the model. As different behaviors are to occur in real case studies, cluster analysis techniques will be of much help on isolating different populations for better determining their characteristics. The use of those varies on the model fitting technique applied later on, but in general terms, having acquired such degree of knowledge about their distribution empowers the successive analyses for more accurate results. Classification analysis techniques can be classified in two large groups: supervised classification ---also known as discriminant analysis--- and non-supervised classification ---also known as clustering.

Discriminant analysis techniques (Arabie et al.,eds. 1996) are useful in those contexts where a finite number of classes are known to exist and to follow a known probabilistic model. Their purpose is twofold:

- Describe the differences amongst occurring classes and define which are the features that best determine or predict the class to which one sample belongs to. These variables are names discriminant variables and the analysis is considered as descriptive.
- Define rules that allow for assigning the observed samples to the occurring classes. This is the predictive analysis.

The purpose of cluster analysis (Gordon, 2010), on the other hand is to identify groups of differentiated behavior within the samples obtained as they are supposed to have been generated by different populations or different states of the generating process. That could be a general definition that in the context of computer vision can be translated to identify different groups of elements pertaining to different objects, images, textures, etc. Predominant clustering techniques can be classified mainly in hierarchical techniques and partitioning techniques.

The fifth step takes advantage of the clustering characterization obtained in step number 4. By considering the differences in the data, a robust model can be obtained. For such a purpose traditional techniques (Grafarend, 2006) for model fitting are especially useful in computer vision, as far as they provide results efficiently requiring small computing efforts and affordable times. This is especially important when analyzing big data sets of images containing several objects per image and describing a considerable amount of key point on each image. Complexity of the algorithms is thus one of the more severe requirements for any application supposed to work on large sets of images.

For smaller sets where more computing power can be spent, neural networks are found especially useful for predictive purposes. Given that multilayer perceptrons can be considered universal function approximators (Hornik et al., 1989) they soon appealed the attention of researcher in need of a tool for model fitting where non-linearities could be

assumed and where second order effects are not necessary neglected for the sake of tractability.

Multilayer perceptron neural networks (MLPNN) consist of a set of units, called neurons, which are grouped into successive layers (Haykin, 2009). Each of these neuron is endowed with a particular behavior, described by their activation function; typically either a hyperbolic, threshold or linear function ---their nature depends on their function on the network, which is typically defined by their location on the different layers. These, the layers, are commonly classified into three categories: input layer, hidden layers and output layer. Neurons in a MLPNN are generally fully connected, which implies that the outputs of the input neurons point towards the inputs of the first hidden layer, the output of the neurons of the first hidden layer point towards the inputs of the second hidden layer or to the inputs of the output layer if the MLPNN is only defined with one hidden layer.

MLPNN are trained with samples from the data set in an iterative manner until some stopping a criterion is applied. The choice for the stopping criterion has an impact on the capability of the neural network to generalize well the learned patterns when new data is entered.

2.4 Hough transform for object recognition

The Hough transform (Dattner, 2009) is a popular technique in the computer vision area. Amongst its numerous advantages robustness can be highlighted. Moreover, it is a fairly efficient algorithm suitable for situations like this where a large set of pictures might be involved. Though originally defined for identifying simple shapes in images, like lines and circles, its use has been extended to more general shapes, allowing for detection of multiple instances of an object that might even be partially occluded. The complexity relies on the number of parameter chosen to represent the complex shape, with search times increasing in exponential order. Fortunately, the Hough transform can be easily implemented on parallel computing systems as each image point can be treated independently (Illingworth et al., 1988) using more than one processing unit.

Essentially, the Hough transform converts sets of points in an image to a parameter space. Thus, two points can be represented in the parameter space as a point in a two dimensional space whose axes represent the two parameters needed to define the line over those two original points in the image. Similarly, the points of circles, ellipses and parabolas in the image can be transformed to points in a new space of parameters. Shape parameterization extends this idea further by means of high-dimensional parameterization that can be decomposed into smaller sets of parameter to determine sequentially.

One of the complexities that appear when using the Hough transform is that for a single image, many points can be chosen to belong to a single line, and thus many lines can be adjusted with the whole data set. Model fitting comes to the rescue in order to choose the best model to use. Additionally, by using the Hough transform a voting scheme can be adopted (Illingworth et al., 1988) which is one of the most common manners of applying the algorithm.

3. Experiments and results

3.1. Dataset

For the purpose of our goal, retrieving query objects from a dataset containing child pornography, we have created our own dataset. It is composed of 614 frames of 640x480 pixels that come from 3 videos recorded under different conditions. All videos were recorded in different bedrooms with different distributions, illumination, textures, etc., making the object

retrieval a challenging task. Nevertheless some objects are present in all videos such as two toy cars, some clothespins, a stuffed bee, some pens, some cups or a child book together with a big doll. The doll is usually the principal actor in the videos and helps us to simulate partial occlusions of the objects and a more realistic scenario. Although these objects are present in every video, they do not appear in every frame. As query objects we have used the book, the blue and yellow car, and the pink, blue and green clothespin shown in Figure 2. The total number of query objects present among the 614 frames that we used in this paper can be found in Table 1.

Table 1. Number of objects in the dataset

Object	Book	Blue car	Yellow car	Pink clothespin	Blue clothespin	Green clothespin
Number of objects	115	102	138	125	92	42

Figure 2: Regions of interest of the query objects.

Book | Blue car | Yellow car | Pink clothespin | Blue clothespin | Green clothespin

The dataset is available at http:// pitia.unileon.es/varp/galleries.

3.2. Results and discussion

When dealing with object retrieval, it is important that the retrieved images are ranked according to their relevance to the query object instead of just being returned as a set. The most relevant hits must be in the top few images returned for a query. Recall and precision are measures for the entire hitlist and do not account for the quality of ranking the hits in the hitlist. Relevance ranking can be measured by computing precision at different cut-off points, this is technically called Precision at n or P@n.

Let *h[i]* be the ith hit in the hitlist and let *rel[i]* be 1 if *h[i]* is relevant and 0 otherwise. Then precision at hit *n* is:

$$P@n = \sum_{k=1..n} rel[k] \ / \ n \quad (1)$$

In order to analyze the retrieval performance of the query objects, we have compared results when relying on the match with the smaller distance and when verifying the location, scale and orientation of the matched keypoints.

In the first case, after matching individual features of the ROI to the features of an image of the dataset using a fast nearest-neighbor algorithm (Lowe, 2004), we chose the match with the smaller distance. Since the smaller distance is considered, this pair of keypoints is the most similar one among all pairs of keypoints matched with the nearest-neighbor algorithm and therefore this match has the highest probability of being correct. We use the distance of

the closest pair of keypoints as a measure of the similarity (distance) between the ROI and the consulted image. The ranking of the retrieved images will be done sorting the images by this distance in ascending order. We will refer to this case as “no clustering”.

In the second case, after matching individual features to a database of features from known objects using a fast nearest-neighbor algorithm, we use a Hough transform to identify clusters belonging to a single object, and finally we perform verification through least-squares solution for consistent pose parameters as suggested by Lowe (2004). Each of SIFT keypoints specifies 4 parameters: 2D location, scale, and orientation, and each matched keypoint in the database has a record of the keypoint’s parameters relative to the training image in which it was found. Therefore, we can create a Hough transform entry predicting the model location, orientation, and scale from the match hypothesis. Lowe’s clustering uses broad bin sizes of 30 degrees for orientation, a factor of 2 for scale, and 0.25 times the maximum projected training image dimension (using the predicted scale) for location. Here we present results with different configurations of the parameters. In this way, half and quarter clustering settings use 60 and 90 degrees for orientation, factor of 4 and 6 for scale, and 0.5 and 0.75 times the maximum projected training image dimension for location respectively. We computed the average of the distances of the matches of every cluster in an image and used the maximum average to rank the retrieved images.

Results for the four clustering types: no clustering, quarter clustering, half clustering and Lowe’s clustering showing the precision at hit 5, 10 and 20 can be seen in Table 2.

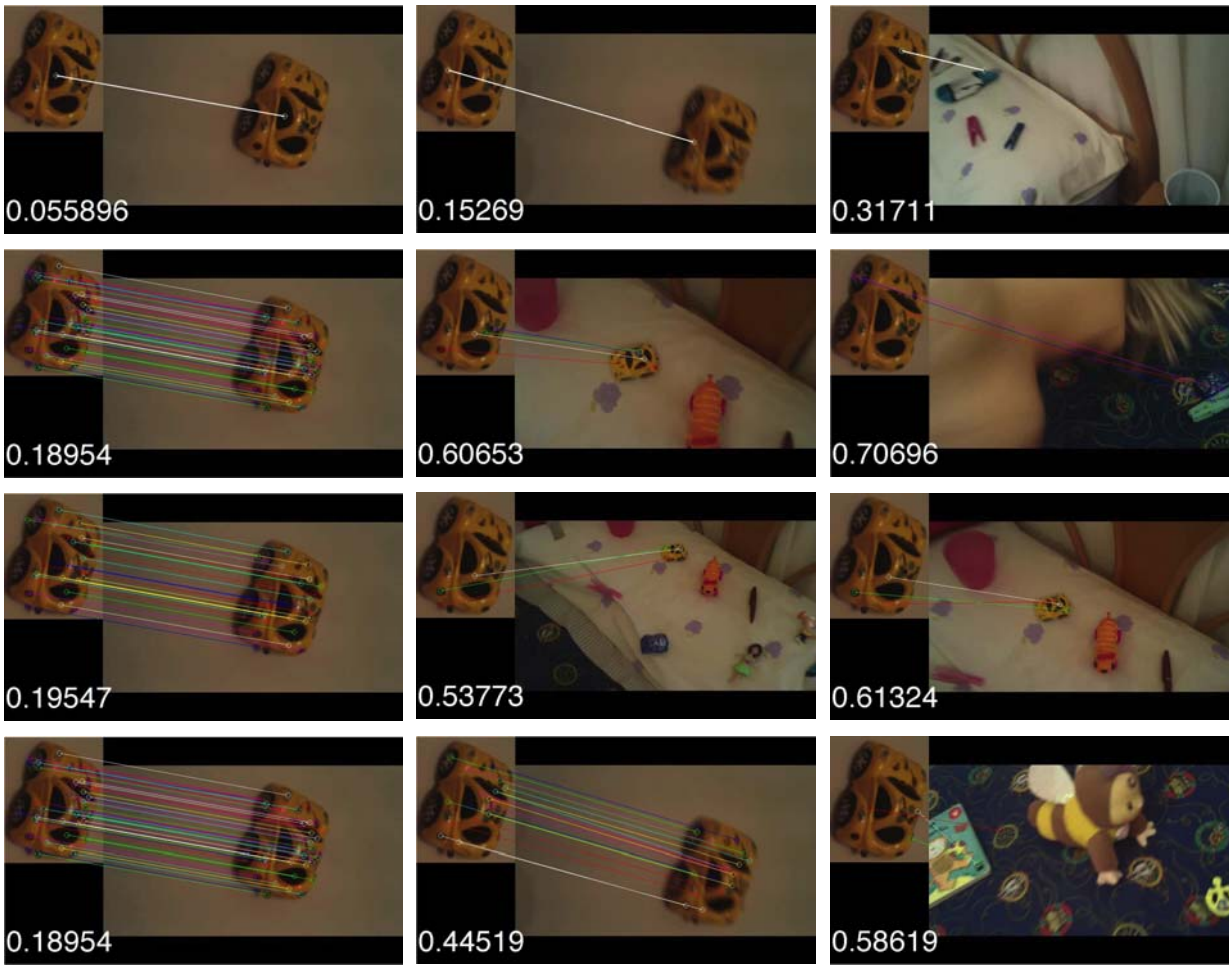
Since SIFT is not invariant to color and the shape of the cars and the clothespins are very similar, it can be assumed that there are three different kinds of objects: book, cars and clothespins. Examples of the second, fifth and tenth hit in the hitlist for the yellow car with the different clustering approaches can be seen in Figure 3.

Table 2: Precision at 5, 10 and 20 of the query objects using different clustering parameters.

	Book			Blue car			Yellow car		
	P@5	P@10	P@20	P@5	P@10	P@20	P@5	P@10	P@20
No clustering	**1**	**1**	**1**	**1**	**1**	**0.75**	**1**	0.9	**0.75**
Quarter clustering	**1**	**1**	0.5	**1**	**1**	0.6	**1**	0.8	0.45
Half clustering	**1**	0.9	0.8	0.2	0.3	0.35	**1**	**1**	0.7
Lowe’s clustering	0.8	0.7	0.7	0.6	0.4	0.3	**1**	0.8	0.7
	Clothespin Pink			Clothespin Blue			Clothespin Green		
	P@5	P@10	P@20	P@5	P@10	P@20	P@5	P@10	P@20
No clustering	**0.8**	**0.4**	**0.25**	**1**	**0.7**	**0.35**	**1**	0.5	0.3
Quarter clustering	0.2	0.1	0.05	0.4	0.2	0.1	0.2	0.1	0.05
Half clustering	0.2	0.3	**0.25**	0.6	0.3	0.15	0.6	0.3	0.15
Lowe’s clustering	0.2	0.3	**0.25**	0.8	0.4	0.2	**1**	**0.6**	**0.35**

Note: Lowe’s clustering uses 30 degrees for orientation, a factor of 2 for scale, and 0.25 times the maximum projected training image dimension (using the predicted scale) for location. Half and quarter clustering settings use 60 and 90 degrees for orientation, factor of 4 and 6 for scale, and 0.5 and 0.75 times the maximum projected training image dimension for location respectively.

Figure 3: Second, fifth and tenth hits using different clustering parameters for the yellow car.



Note: Rows: No clustering, quarter clustering, half clustering and Lowe's clustering. Columns: second, fifth and tenth hit in the hitlist. The number indicates the distance of the match or the avarege distances of the matches.

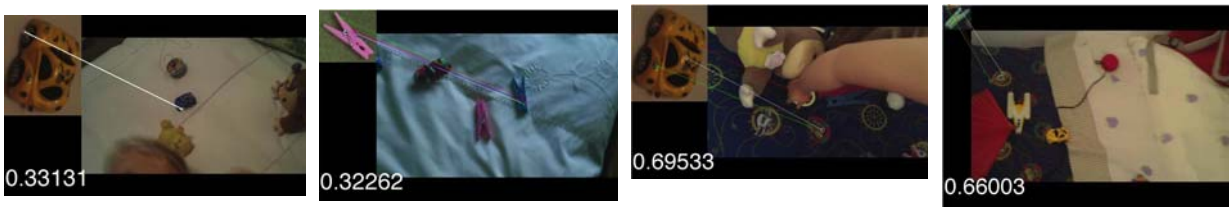
The ROI of the book is a 350x334 image with well-defined corners that should be quite easy to detect and match among the images of the dataset. Table 2 shows a perfect retrieval for the first 20 matches when no clustering approach is used. Results get worse as a clustering more similar to Lowe's one is performed. In fact, precision at 5 is 1 (all 5 first hits were relevant) except for Lowe's clustering in which one hit is not relevant. Moreover, using the no clustering approach retrieves the 51 first images correctly, considering that there are 115 images containing the book in all 614 dataset images this is a very promising result.

For the car queries we used a ROI of 285x258 for the blue one and a ROI of 208x265 for the yellow one. Although SIFT method computes the descriptors using gray level images, there are small differences in shape and patterns between the two cars that can help the algorithm to distinguish them. The blue car retrieval is perfect for the first 5 and 10 hits either using no clustering or quarter clustering approaches. Precision at 20 increases when less restrictive clustering is used, being P@20 equals to 0.75 with a no clustering approach. In regard to the yellow car, we can observe that the first 5 images are well retrieved no matter the clustering configuration used. Precision at 10 is 1 for half clustering approach followed whereas precision at 20 is better for no clustering configuration achieving a value of 0.75.

As to the clothespins, we used ROIs of 146x132, 85x145 and 68x59 for the pink, blue and green one respectively. Clothespins present a more difficult task since less distinctive keypoints are present. Most of the keypoints are found in the metal wire, near the holes or in the outlines. Precision at 20 only reaches 0.25, 0.35 and 0.35 respectively. No clustering configuration is proved to work better for retrieving the pink and blue clothespins while Lowe's clustering performs better for the green clothespin.

In order to show how similar objects with different colors can lead to mismatches, in Figure 5 (first two images) we present examples of misclassified query objects. Only two mismatches among different cars appeared over all our experiments for the first 20 hits of the hitlist but up to 10 in the case of the clothespins. Background is also another source of mismatches. The pattern duvet of one of the settings lead to many non relevant but distinctive keypoints that locally described can look similar to other objects. We have found out that some of the patterns of the duvet are similar to the patterns of the yellow car. Figure 4 (second two images) shows some examples of this fact.

Figure 4: Mismatches produces by similar objects (first two) and pattern duvet (second two).



Finally, we would like to make a comment about the number of hits found with each configuration. No clustering approach retrieves all images of the dataset in the hitlist unless there are some images with very few keypoints, which is unusual. On the contrary when considering clusters to vote for the same object, more images in the dataset do not produce a hit. Quarter clustering configuration obtains the fewest hits, followed by half clustering and Lowe's clustering approaches. For example, the number of hits for the green clothespin was 614, 18, 4 and 1 with no clustering, Lowe's clustering, half clustering and quarter clustering respectively. This becomes a restriction in our application field since we prefer false positives rather than false negatives in order to not loose information of possible child abuses sceneries.

4. Conclusions

In this paper we have presented and compared four different clustering approaches to carry out the retrieving of different query objects in a 614-image dataset. After applying SIFT descriptors and a fast nearest-neighbors matching, Hough transform with three configurations of the parameters and least squares refinement were used to make clusters vote for the object pose, orientation and scale. Also, we evaluated a simpler case that relies on the closest match of the nearest-neighbors matching. Results are not conclusive but show some interesting facts. Most of the times, no using a clustering outperforms the rest of the configurations. One reason we have observed is that the closest match between a ROI and an image can be lost because there are not at least other three points agreeing with the objects pose, scale and orientation. Another thing to be considered is the fact that fewer images are retrieved when using a softer configuration of the clustering parameters. Besides, similar objects and cluttered background have led to quite a few misclassifications. All in all, this technique could help to retrieve images from child pornographic datasets, which is a huge improvement since it is a task usually manually performed nowadays. In future work we

will perform evaluation with a bigger number of query objects and apply different techniques of clustering such as RANSAC.

Acknowledgements

This work has been supported by grant DPI2012-36166 from the Spanish Government and by the Advisory System Against Sexual Exploitation of Children (ASASEC) European Union project with reference HOME/2010/ISEC/AG/043

References

Alahi, A., Ortiz, R., & Vandergheynst, P. (2012). FREAK: Fast Retina Keypoint. *IEEE Conference on Computer Vision and Pattern Recognition*.

Arabie, P., Lawrence, J. and Geert De Soete, A., eds. (1996) *Clustering and classification*. World Scientific Publ.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. *European Conference on Computer Vision*.

Calonder, M., Lepetit, V., Strecha,, C. & Fua, P. (2010). Brief: Binary robust independent elementary features. *European Conference on Computer Vision (ECCV)*, 778–792.

Dattner, I. (2009) Statistical properties of the Hough transform estimator in the presence of measurement errors. *Journal of Multivariate Analysis (100),* no. 1, 112-125. Doi: 10.1016/j.jmva.2008.03.005

Gordon, A.D. (2010) *Classification, Monographs on Statistics and Applied Probability, 82.* Chapman and Hall.

Grafarend, E. (2006) Linear and nonlinear models: fixed effects, random effects and mixed models. Book News, Inc. Portland.

Han, J. & Kamber, M. (2006) Data Mining, concepts and techniques. Elsevier.

Haykin, S.S. (2009) Neural Networks and learning machines. Prentice Hall.

Hornik, K., Stinchcombe, M., White, H. (1989) Multilayer feedforward networks are universal approximators. *Neural networks (2),* no. 5, 359-366, doi:10.1016/0893-6080(89)90020-8

Illingworth, J. & Kittler, J. (1988) A survey of the Hough transform. *Computer Vision, Graphics and Image Processing (44)*, 87-116 doi: 10.1016/S0734-189X(88)80033-1

Leutenegger, S., Chli, M., & Siegwart, R. (2011). Brisk: Binary robust invariant scalable keypoints. *IEEE International Conference on Computer Vision (ICCV)*.

Lowe, D (1999). Object Recognition from Local Scale-Invariant Features. In *Seventh Int'l Conference on Computer Vision* (pp. 1150- 1157).

Lowe, D. (2004). Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision, 2*, 91–111.

Mair, E., Hager, G. D., Burschka, D., Suppa, M. and Hirzinger, G. (2010). Adaptive and generic corner detection based on the accelerated segment test. In *Proceedings of the Eu- ropean Conference on Computer Vision (ECCV).*

Ozuysal, M., Calonder, M., Lepetit, V., & Fua, P. (2010). Fast keypoint recognition using random ferns. *Pattern Analysis and Machine Intelligence, 32*, 448–461.

Rosten, E. and Drummond, T. (2006). Machine learning for high- speed corner detection. In *European Conference on Computer Vision (ECCV)*, (1)

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). Orb: an efficient alternative to sift or surf. *International Conference on Computer Vision*.