Fast Human Activity Recognition in Lifelogging

Stefan Terziyski, Rami Albatal, and Cathal Gurrin
{stefan.terziyski, rami.albatal, cathal.gurrin}@insight-centre.org

Insight Centre for Data Analytics,
Dublin City University, Dublin, Ireland

Abstract. This paper addresses the problem of fast Human Activity Recognition (HAR) in visual lifelogging. We identify the importance of visual features related to HAR and we specifically evaluate the HAR discrimination potential of Colour Histograms and Histogram of Oriented Gradients. In our evaluation we show that colour can be a low-cost and effective means of low-cost HAR when performing single-user classification. It is also noted that, while much more efficient, global image descriptors perform as well or better than local descriptors in our HAR experiments. We believe that both of these findings are due to the fact that a user's lifelog is rich in reoccurring scenes and environments.

1 Introduction

Recent technological development in personal and ubiquitous computing, data storage and computational power has provided the environment for lifelogging to become a normative activity [1]. At the same time, Human Activity Recognition (HAR), a well-established research field is receiving increasing attention, however much of this attention has been directed towards using fixed cameras observing people, rather than from the point of view of the wearer. HAR from lifeloggers' point of view visual context has been described by Steve Mann as "sousveillance" (different than the classic surveillance concept) in [2]. In sousveillance we do not get to observe what the human or lifelogger is doing, but rather the view that the lifelogger is seeing. This provides a visual context of the user and this visual context is a source for HAR.

The purpose of this research effort is to examine the feasibility of fully-automatic, low-cost HAR in lifelogging. When considering HAR in lifelogging we mean the process of identifying what activity the wearer is performing at the moment of capture. While HAR may be performed at any point in the lifecycle of lifelogging data (capture time, processing time, access and feedback time), our key interest here is in supporting real-time low-cost HAR using visual lifelog data. Consider the following scenario, a Google Glass wearer, upon the advice of his doctor, is interested in identifying and logging his dietary habits. Given current battery considerations, continual capture and upload to a server for analysis is not feasible. A HAR algorithm, successfully running on his device would be able to trigger capture of a series of detailed photos upon detection of and eating event.

Another scenario is where a lifelogger is interested in storing a record of his life activities, but due to privacy considerations with wearable cameras, he does not want to transmit and store wearable camera photos. Such considerations motivate this paper.

The ability to capture visual user context in real-time would be an enabling technology for many applications. Considerations of battery life and network-transport delay means that continual transmission of visual content to a server for analysis and feedback is far from ideal. In this work, we illustrate how low-cost HAR can be achieved on a per-user basis and we evaluate this by means of an experimental evaluation on data gathered from 5 users over 14 days of visual lifelog data. It is our conjecture that a low-cost implementation of HAR would be more applicable to on-device deployment than more conventional techniques such as SIFT. We show that the colour distribution combined with the global image gradient intensity performs well on low resolution images for a HAR task for an individual user.

The rest of the paper is as follows: in 2 we outline previous work done on activity recognition. In section 3 we provide a description of the activities, the visual data and the features extraced for it. Subsequently in section 4 we describe our evaluation method, then in section 5 the results we have obtained and in section 6 what conclusions we can draw from this initial exploration.

2 Related works

Lifelogging represents a phenomenon whereby individuals can digitally record their own daily lives in varying amounts of detail and for a variety of purposes. Although there are many definitions in the literature, we define lifelogging to be *a form of pervasive computing which utilises software and sensors to generate a permanent, private and unified multimedia record of the totality of an individual's life experience and makes it available in a secure and pervasive manner.* A key aspect of this definition is that the lifelog should archive the totality of an individual's experiences, outlined in Bell and Gemmels' vision of total capture [3]. This means that lifelogging will generate a rich set of multimedia data, gathered from wearable sensors on the lifelogger.

One early attempt in activity recognition in lifelogging is by Doherty, et al. in [4], where the authors aimed at developing a trait interpreter tool in lifelogging data. Some of the target traits were activities, such as shopping and using mobile phone. An average crossover accuracy of 61% was reported. The crossover is between system output and lifelogger self-report. The authors used MPEG-7 features, namely ScalableColor and ColorLayout.

There is a research direction within the domain of HAR from visual lifelog data, where certain objects are detected and the co-occurrence of these within visual images forms the basis of activity recognition, as by Wang and Smeaton in [5], where they report an average F1-score over all activities, used in their experiments, of 90% with a baseline object detection accuracy of 65% on average. A system with such classification framework would be difficult run in real-time context, because it is necessary to determine what objects are needed to detect in order to determine the activities for a particular user. In practice human assistance would be needed for that, hence the system would not be fully automatic.

Another approach is the work by Hamm et al. in [6], where global image descriptors are combined with other sensory data to form what the authors call a multi-sensory bag-of-words. The work produces an average F1-score for the best classifier of 92%, however this is averaged from the F1-scores of the trained classifiers per activity. The authors do not propose a model, where they fuse all of the per activity classifiers to form a final activity classification model. If, for example, two activities were mutually exclusive then the system should be able to choose only one activity, but not both. The sensors that they use are: digital camera, accelerometer, audio and GPS. The authors have shown that fusing different sensors improves the accuracy and also show that discriminative models tend to perform better on global descriptors. The visual features used in [6] are colour histograms in HSV-space.

In this work we aim to provide for a fully automatic, low-cost and fast classification model for HAR from visual lifelog data that has potential to be deployed on the lifelogging device. We achieve that by examining a more direct relationship between the visual context and the activity being performed, thus rendering our classification process automatic. We employ computationally inexpensive visual features operating over low resolution images.

3 HAR Technique

Our HAR technique starts with visual feature extraction from image data. The chosen feature(s) (which can be any or a combination of Colour Histogram, HOG, SIFT) are the one(s) that yielded the best results. Different parameters (such as image resolution, colour bins, number of gradient orientations, etc.) would also be tuned at this stage. Then the feature set is used for training a per-activity and per-user SVM classifier (with radial-basis kernel). We aim to improve the performance of those classifiers. Our experimental data set is divided into training and evaluation subsets as it is explained later on in section 4.

In developing the HAR technique, we identified five research questions that we needed to address. Firstly, the experiments that we set up were to determine whether the amount of training examples affects the classification performance. This was achieved by comparing the F1-scores per activity for two data sets, where the second one contained an additional training examples for one activity. Secondly, given the same configuration, we examined whether this affected the classification performance for the rest of the activities. Thirdly, we examined various combinations of features namely colour with texture and local with global features. Fourthly, we carried out tests to see what feature scaling techniques would be most appropriate. Finally we compare the computational efficiency of extractions of the different features.

3.1 Human Activities

In selecting a set of human activities for classification, we turned to the work of Kahneman, et al. in [7], in which the 15 most enjoyable daily activities of people were identified by means of a large-scale survey. Of the 15 activities identified, we have chosen 9 for this work. We purposely excluded 5 activities, namely 'intimate relations', 'relaxing', 'pray/worship/meditate', 'napping' and 'taking care of children', because we consider those to be private or intimate and beyond the scope of our current research. One final activity, 'exercising', is excluded because it was not actually present in the lifelog of the any of 5 users. This is most probably due to the difficulties of wearing a lanyard mounted camera during exercise.

3.2 Visual Data and Annotations

As mentioned above, this paper focuses on processing the visual input. Those are lifelog images taken from an OMG Autographer wearable camera[1] which is a fish-eye camera that can be worn on a lanyard around the neck, or clipped onto clothing. In normal use, it is oriented with the viewpoint of the wearer. We had the camera configured to capture an image approximately every 20-30 seconds.

The data set consists of 41,397 images from 5 users, with 14 days of lifelog data per user. The experiments were carried out mainly on one user's data which consists of 4,315 images. The rest of the data from 4 other users with a total of 37,082 images was used for validating the cross-user generalisation of the selected features. It is to be noted that the lifelogging was not continuous and some gaps may occur, in terms of consecutive days. See Fig. 1 for examples of the visual lifelog data that we employed. The visual repetition of some human activities are clearly visible.

We note the unbalanced nature of lifelog data, due to the naturally varying lifestyle across the users. Refer to Table 1 for a detailed distribution. In Table 1 'N/A' stands for Not Available.

Table 1. Activity distributions among users' lifelog data in number of images

Activity	**User 1**	**User 2**	**User 3**	**User 4**	**User 5**
1.Commuting	1317	2068	2937	373	2566
2.Computer	2140	5223	3969	4961	3337
3.Eating	59	505	538	558	677
4.Housework	63	205	90	57	111
5.On the Phone	23	1087	436	N/A	227
6.Preparing Food	33	130	144	989	12
7.Shopping	N/A	98	228	13	408
8.Socialising	680	845	487	525	2800
9.Watching TV	N/A	192	386	N/A	N/A

[1] www.autographer.com

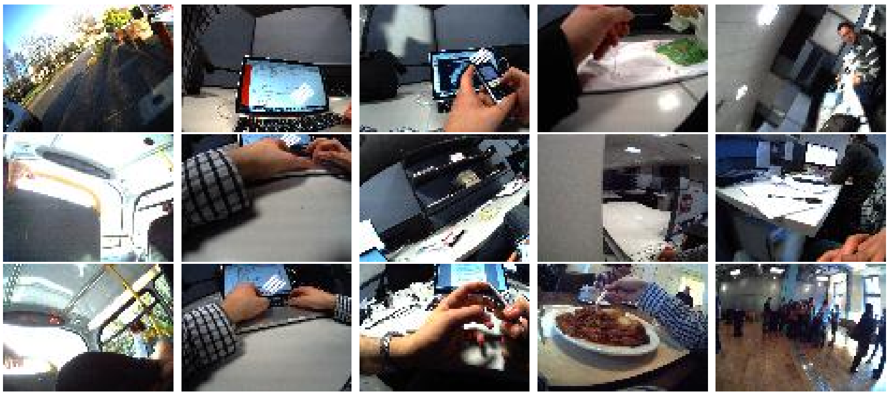

Fig. 1. Example activities showing the similar visual context of the activities. This is also shows that some activities occur in the same environment, which affects the classification process. From left column to right column: commuting, computer, on the phone, eating, socialising.

3.3 Visual Features

For fast image processing, the resolution of the lifelog images was reduced to 91x68 pixels keeping the original aspect ratio. We chose two features, namely colour histogram and Histogram of Oriented Gradients (HOG) by Dalal and Triggs in [8] and we examine the importance of colour and texture and the importance of local and global descriptors and their combination. Colour histogram is a global image descriptor and provides one colour distribution per image. The latter feature (HOG) is a local descriptor and provides texture information per image.

Colour Histogram. For each image colour histograms were extracted, where the combination of the RGB values are preserved. We extracted Colour Histograms at 16 bins, 8 bins and 4 bins per each channel. More detailed colour information such as 32, 64, 128 and 256 bins per channel, may prove more valuable, however the feature vectors would be larger, which could have a consequential impact on the computational time of the SVM classification process. Hence we compared 4, 8 and 16 bins per channel, giving 64, 512 and 4096 bins respectively. All of them have comparable performance and we can report that 512 bins and 4096 bins give similar performance on the test that we have performed, therefore we chose the 512 bin histogram due to the considerable reduction in computation time, which is due to the time complexity of the SVM. We have subsequently used the 512 bin RGB histogram for all of the experimentation in this paper, since the colour precision at this point proves to be sufficient.

Histogram of Oriented Gradients. HOG features were extracted. The format is: 9 orientation bins; 8 by 8 pixels per cell; 2 by 2 cells per block, hence

16 by 16 pixels per block. Hence a feature vector of length 2,520 is produced, where 280 cell responses are captured, giving 9 orientations each. The amount of cell responses is derived from the per-block normalisations, however the blocks are overlapping, as suggested in [8]. The parameters were also chosen as per [8], since those have shown to give the best results. The authors' aim in [8] was to improve object recognition from a detection window and they suggest that this gives comparative results to the state-of-the-art object classifiers. We have chosen HOG as it describes a scene by describing the gradient intensities in each direction, per cell and those normalised per neighbouring cells. We will henceforth refer to the original HOG as 'local' HOG.

We altered HOG's locality by constructing a 'global' HOG and also evaluated that. The evaluation of the 'global' HOG is done in order to evaluate whether the locality of a feature affects the classifier. Consider the following scenario: given several 'computer' scenes, the computer may appear at different locations in the scenes, which would affect the local HOG, but would not affect the global HOG. This hypothesis was evaluated by the introduction of 'global' HOG.

We did so by modifying the parameters, so that the cell becomes 91 by 68 pixels and there is only 1 by 1 cell per block, thus we allow the algorithm to only compute the gradient intensity for the entire image, instead of breaking it down into cells and blocks. This can be seen as global texture descriptor, similar to but not the same as an edge orientation histogram. In the latter only edges are counted per direction.

Considering Scale Invariant Feature Transform Scale Invariant Feature Transform (SIFT) could be considered to be a widely deployed feature extraction approach. Hence we employed SIFT and we benchmarked the performance of SIFT when compared to the features employed in this paper. We used dense sampling on every 6 pixels and OpponentSIFT for a descriptor as suggested and implemented by Koen van de Sande in [9]. We used the codebook model to construct a codeword histogram per image and evaluated that with a RBF SVM. The resulting descriptor of an image is a normalised histogram of all the visual words encountered in the image, binned into their respective codewords via clustering. We chose a codebook of size 512, which reported comparable results to those obtained by RGB and HOG features when using the full sized image, but lower performance when using low resolution images, where many visual details might be lost.

For the computing performance comparison, colour histogram and SIFT were employed as executable binary files, implemented in C/C++, and were ran on the same machine over the same set of images, performing the same input-output operations. We report that the average extraction time of SIFT descriptor from one image is 0.05s, as per the implementation provided [9], whereas using an RGB histogram extraction, as provided by OpenCV takes about 0.00037s and a binary executable implementation of HOG extracts the feature in 0.00032s. Thus the total extraction time of HOG and RGB histograms together is 0.0069, which make both of these features many times quicker to extract than SIFT. We do not consider the codebook creation also, even though it is a lengthy process, since it is only needed once, whereas the feature extraction process per image is required for all new input.

For this reason, and given the focus of this research on fast and efficient local processing that could be applied to a relatively low-power device (such as Google Glass), we decided that the overhead of implementing SIFT was too great for our use-case, so we proceed to compare and evaluate the two prior algorithms.

4 Evaluation

Evaluation was carried out via splitting the data set into 80% training examples and 20% examples for evaluation. In this instance the original distribution of data was kept, meaning that for each activity the amount of training examples is 80% and the amount of evaluation examples is 20% out of the total available examples of that activity.

The evaluation metric that we used is F1-score. We obtain the F1-score, as it is expressed in terms of the harmonic mean of precision and recall. However, we work within a classification context, hence both the precision and the recall were obtained in that context. In the classification context, TP refers to True Positives, FP refers to False Positives and FN refers to False Negatives, hence we calculate the F1-score as:

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \end{aligned}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

When performing the classification of the various features, we also used feature scaling, because that improved the performance of all activity classifiers over the same features. Based on a prior evaluation, the feature scaling outlined in equation (1) was selected because it had proven to be the most successful in initial evaluations on colour histograms and HOG.

$$\begin{aligned} x_j &= (x_j - mean)/range \\ x_j &= x_j/max \qquad (1) \\ x_j &= x_j/\sum_X \end{aligned}$$

5 Results

We noted that for the visually similar activities, the number of the positive training examples affected the F1-score, hence the performance of the classifier.

Whereas for activities whose context is more visually dissimilar that was not necessarily the case. By dissimilar we mean that the visual context is different and vice versa for visually similar.

Table 2. F1-scores of the available activities for User 1. Each activity has its own classifier.

Activity	**local HOG**	**RGB Hist**	**global HOG + RGB Hist**
1.Commuting	0.82	0.94	0.95
2.Computer	0.88	0.92	0.93
3.Eating	0.59	0.66	0.77
4.Housework	0.27	0.13	0.35
5.On the Phone	0.40	0.66	0.86
6.Preparing Food	0.67	0.57	0.67
7.Socialising	0.57	0.77	0.77

See Table 2 'local HOG' for results. When using global HOG, we had actually found an F1-score of 0 for all activities except for 'computer', where that was 0.66, so this did not perform well at all.
When combining both the HOG features and the colour histograms, we used both types of HOG. When using the local HOG we had no success whatsoever, in spite of the promising results from both in Table 2 - F1-scores for User 1 above. We got low F1-score throughout the activities. But when we combined the RGB histogram with the global HOG descriptor we got an improvement in the classification performance. The RGB histogram alone gives an average F1-score of 65%, whereas in combination with the global HOG the average F1-score becomes 75%. See Table 2 under 'global HOG + RGB Hist'.

Given the results in Table 2, we can say that colour provides for a greater discriminative potential than texture when comes to HAR. We can also say that a global image descriptor in our case could outperform local ones. We have obtained the best results from using a RGB histogram combined with a global HOG.

With regards to the SVM for the activity classification process, we used a radial-basis-function kernel SVM with hyper-parameters C=1.0 and γ=1.0 for all of the results presented, except where stated otherwise. We present a grid-search optimisation for C being 10 whose exponent ranges from -2 to 2 and γ being 10 whose exponent ranges from 2 to -4. This was used purely for exploratory data analysis in order to understand the relationship between the visual context and performed activity by the wearer. If we observe a high C, that means that those activities have similar visual contexts that are difficult to distinguish between. Respectively for γ where this hyper-parameter is low it means that the data is diverse, whereas with high values it means that those activities have a less diverse visual context.

Given Table 3, we can see that in terms of texture as a local descriptor, activi-

Table 3. C and γ optimisation in the SVM Classifier

	local HOG			**RGB Hist**		
Activity	**F1-score**	**C**	γ	**F1-score**	**C**	γ
1.Commuting	0.89	10	0.1	0.95	1	1
2.Computer	0.92	0.10	0.1	0.93	100	1
3.Eating	0.62	0.01	0.1	0.77	100	1
4.Housework	0.47	0.01	0.1	0.29	0.01	0.1
5.On the Phone	0.66	10	0.1	0.66	0.01	1
6.Preparing Food	0.66	1	0.01	0.80	10	1
7.Socialising	0.58	10	0.1	0.81	10	1

ties 1, 5 and 8 are difficult to distinguish from the rest of the training examples. Whereas in the case of the colour histogram, activities 2, 3, 6 and 8 are difficult to discriminate from the rest, due to the training examples overlapping with each other. All activities' classifiers whose C parameter is above 1 are considered to generalise poorly for new examples.

We validated the importance of colour for activity recognition over the rest of the lifelog data provided by the other four users. See the Table 4 for a more detailed information.

Table 4. F1-scores for the rest of the users on Colour Histograms

	User 2		**User 3**		**User 4**		**User 5**	
Activity	**Qty**	**F1**	**Qty**	**F1**	**Qty**	**F1**	**Qty**	**F1**
1.Commuting	2068	0.86	2937	0.92	373	0.47	2566	0.81
2.Computer	5223	0.92	3969	0.90	4961	0.96	3337	0.94
3.Eating	505	0.51	538	0.53	558	0.70	677	0.84
4.Housework	205	0.57	90	0.09	57	0.74	111	0.57
5.On the Phone	1087	0.68	436	0.63	N/A	N/A	227	0.41
6.Preparing Food	130	0.74	144	0.49	989	0.85	12	0.0
7.Shopping	98	0.5	228	0.60	13	1.0	408	0.73
8.Socialising	845	0.64	487	0.50	525	0.74	2800	0.83
9.Watching TV	192	0.59	386	0.90	N/A	N/A	N/A	N/A

It is our consideration that colour acts as a discriminator between the activities that a user can be engaged in during daily life and based on our experimentation, this has proven to be sufficient to classify the activities within the data of one user. However, the relation between the user's activity and the colour distribution of the visual context is user-dependent. We applied cross-user validation, meaning that the classifier was trained on one user's data and then used for validation on the another user's data. Training on user 3 and validating on user 2 has an average F1-score of 19% for all activities. The rest is as follows:

user 3 on user 4 is 12%; user 3 on user 5 is 18%; user 2 on user 4 is 9%; user 3 on user 5 is 20%.

6 Conclusion

In this paper we examined the importance of texture and colour as well local versus global image descriptors in order to determine the features with the best classification performance and low-cost extraction time. We evaluated the performance of HOG, RGB Histograms and combinations thereof and used SIFT as a benchmark for time tests.

There are several conclusions that we can derive from our results. First, local HOG descriptors do not always provide the best classification performance in our scenario. Although what we named local HOG outperformed significantly what we proposed as global HOG, it (the local HOG) was outperformed by the colour histogram. Second, we have verified the importance of colour on the rest of the lifeloggers' data and we can show that the colour distribution of the visual context of a lifelogger is related to the activity, which the lifelogger is performing. Third, texture information is important, and we can confirm that it affects the performance of the classifier positively as with the global HOG. Lastly, the trained classifiers are user-dependent, therefore classifier trained on one lifelogger cannot used for inference on other lifeloggers.

We also demonstrated how real-time HAR is possible with computational and storage efficient processing, due to the low resolution that still provides good results. Finally, we cannot state that the visual aspect for HAR is alone sufficient. The performance of HAR on low-power devices could potentially be significantly enhanced by the integration of additional sensors, such as accelerometers, location, etc.

Future Work

With regard to limitations of this work, we have noted different luminance conditions in the images and we believe that this may lead to misclassification. Considering alternative colour models, such as HSV, is a further step, where the idea is to neutralise any shadows that may appear in the visual context due to different luminance conditions in the lifelog.

We also recognise the need for a larger sample of people in order to validate our findings. We also note the necessity for more lifestyle activities. This is due to a possible direction of the research into fully characterising the day of a lifelogger as opposed to identifying occasional scenes where the lifelogger is performing a certain activity.

A further direction of the research may go into a multisensory approach as that will inevitably increase the classification performance. We also confirm the need, as suggested by previous research that an ontology for an activity needs to be defined as that may have a direct impact on the classification framework and hence the classification performance.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289.

Third-party libraries were used for feature extraction, specifically [10] for HOG extraction, [11] for SVM classification and [9], [12] for SIFT comparison.

References

[1] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. “LifeLogging: Personal Big Data”. In: *Foundations and Trends® in Information Retrieval* 8 (2014), pp. 1–125.

[2] Steve Mann. “Continuous lifelong capture of personal experience with EyeTap”. In: *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences.* New York, New York, USA: ACM, 2004, pp. 1–21.

[3] Gordon Bell and Jim Gemmell. *Total Recall: How the E-Memory Revolution Will Change Everything.* Dutton, 2009.

[4] Caprani N. Conaire C. Kalnikaite V.-Gurrin C. Smeaton A.F. Doherty A.R. and N.E. O’Connor. “Passively Recognising Human Activities Through Lifelogging”. In: *Comput. Hum. Behav.* 27.5 (2011), pp. 1948–1958.

[5] Peng Wang and Alan F. Smeaton. “Using Visual Lifelogs to Automatically Characterize Everyday Activities”. In: *Inf. Sci.* 230 (2013), pp. 147–161.

[6] Stone B. Belkin M. Hamm J. and S. Dennis. “Automatic Annotation of Daily Activity from Smartphone-Based Multisensory Streams”. In: *Mobile Computing, Applications, and Services.* Vol. 110. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer Berlin Heidelberg, 2013, pp. 328–342.

[7] Krueger A.B. Schkade D.A. Schwarz N. Kahneman D. and A.A. Stone. “A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method”. In: *Science* 306.5702 (2004), pp. 1776–1780.

[8] N. Dalal and B. Triggs. “Histograms of oriented gradients for human detection”. In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on.* Vol. 1. 2005, 886–893 vol. 1.

[9] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. “Empowering Visual Categorization with the GPU”. In: *IEEE Transactions on Multimedia* 13.1 (2011), pp. 60–70.

[10] Schönberger J. L. Nunez-Iglesias J. Boulogne F. Warner J. D. Yager N. Gouillart E. van der Walt S. and T. Yu. “scikit-image: Image processing in Python”. In: *PeerJ* 2 (June 2014), e453.

[11] Varoquaux G. Gramfort A. Michel V.-Thirion B. Grisel O. Blondel M. Prettenhofer P. Weiss R. Dubourg V. Vanderplas J. Passos A. Cournapeau D. Brucher M. Perrot M. Pedregosa F. and E. Duchesnay. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[12] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. “Evaluating Color Descriptors for Object and Scene Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.9 (2010), pp. 1582–1596.