

VISUAL SUMMARY OF EGOCENTRIC PHOTOSTREAMS BY REPRESENTATIVE KEYFRAMES

Marc Bolaños, Estefanía Talavera and Petia Radeva

Universitat de Barcelona
{marc.bolanos,etalavera,petia.radeva}@ub.edu

Ricard Mestre and Xavier Giró-i-Nieto*

Universitat Politècnica de Catalunya
xavier.giro@upc.edu

ABSTRACT

Building a visual summary from an egocentric photostream captured by a lifelogging wearable camera is of high interest for different applications (e.g. memory reinforcement). In this paper, we propose a new summarization method based on keyframes selection that uses visual features extracted by means of a convolutional neural network. Our method applies an unsupervised clustering for dividing the photostreams into events, and finally extracts the most relevant keyframe for each event. We assess the results by applying a blind-taste test on a group of 20 people who assessed the quality of the summaries.

Index Terms— egocentric, lifelogging, summarization, keyframes

1. INTRODUCTION

Lifelogging devices offer the possibility to record a rich set of data about the daily life of a person. A good example of this are wearable cameras, that are able to capture images from an egocentric point of view, continuously and during long periods of time. The acquired set of images comes in two formats depending on the device used: 1) high-temporal resolution videos, which usually produce more than 30fps and capture a lot of dynamical information, but they are only capable of storing some hours of data, or 2) low-temporal resolution photostreams, which usually produce only 1 or 2 fpm, but are able to capture events that happen during a whole day (having around 16 hours of autonomy).

Being able to automatically analyze and understand the large amount of visual information provided by these devices would be very useful for developing a wide range of applications. Some examples could be building a nutrition diary based on what, where and in which conditions the user eats for keeping track of any possible unhealthy habit, or providing an automatic summary of the whole day of the user for

offering a memory aid to mild cognitive impairment (MCI) patients by reactivating their memory capabilities [1].

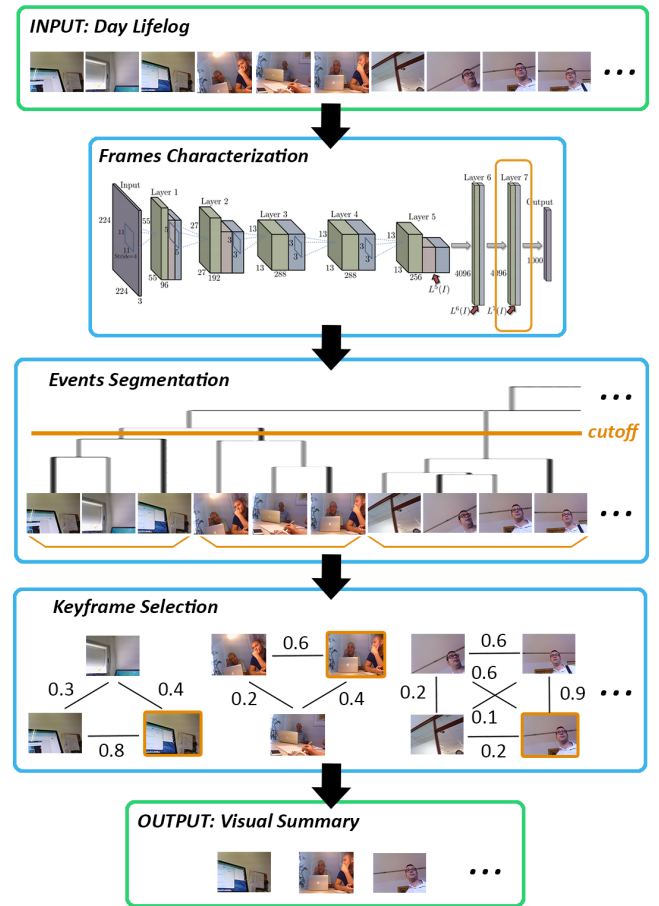


Fig. 1. Scheme of the proposed visual summarization.

In our work, we focus on automatic extraction of a good summary that can be used as a memory aid for MCI patients. Usually, these patients suffer neuron degradation that generates them problems to recognize familiar people, objects and places [2]. Hence, the visual summary, automatically extracted, should be clear and informative enough to recall the daily activity with a simple glimpse.

In order to take into account our ultimate goal, we pro-

*This work has been developed in the framework of the project BigGraph TEC2013-43935-R, funded by the Spanish Ministerio de Economía y Competitividad and the European Regional Development Fund (ERDF). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GeoForce GTX Titan Z used in this work.

pose an approach that starts by extracting a set of features for frames characterization by means of a convolutional neural network. These visual descriptors are used to segment events by running an agglomerative clustering, which is post-processed to guarantee a temporal coherency (similar to [3]). Finally, a representative keyframe for each event is selected using the Random Walk [4] or Minimum Distance [5] algorithms. The overall scheme is depicted in Figure 1.

This paper is structured as follows. Section 2 overviews previous work for event segmentation and summarization in the field of egocentric video. Our approach is described in Section 3 and its quantitative and qualitative evaluation in Section 4. Finally, Section 5 draws the final conclusions and outlines our future work.

2. RELATED WORK

The two main problems addressed in this paper, event segmentation and summarization, have been addressed in related egocentric data works, as presented in this section.

2.1. Egocentric event segmentation

Most existing techniques agree that the first step for a summary construction is a shot- or event-based segmentation of the photostream or video. Lu and Grauman in [6] and Bolaños et al. in [7] both propose event segmentation that relies on motion information, colour and blurriness, integrated in an energy-minimization technique. The result is final event segmentation that is able to capture the different motion-related events that the user experiences. In the former approach [6], the authors use high-temporal videos and an optical flow descriptor for characterizing the neighbouring frames. In the latter one [7], instead of working with low-temporal data, a SIFT-flow descriptor is used, as it is more robust for capturing long-term motion relationships. Poleg et al. [8] also propose motion-based segmentation, but they use a new method of Cumulative Displacement Curves for describing the motion between neighbouring video frames. The proposed solution is able to focus on the forward user movement and removes the noise of the head motion produced by head-mounted wearable cameras. Other methods have been proposed using low-level sensor features like the work in [9] that splits low-temporal resolution lifelogs in events. Lin and Hauptmann [3] also propose a simple approach based on using colour features in a Time-Constrained K-Means clustering algorithm for keeping temporal coherence. In [10], Talavera et al. design a segmentation framework also based on an energy minimization framework. In this case, the authors offer the possibility to integrate different clustering and segmentation methods, offering more robust results.

2.2. Egocentric summarization

Focusing on the summarization of lifelog data after event segmentation, there are two basic research directions, both of them aiming at removing those data, which are redundant or low-informative. In the case of video recordings, it is a common practice the select a subset of video segments to create a video summary. On the other hand, when working with devices that take single pictures at a low frame rate, the problem is usually tackled by selecting the most representative keyframes. The most relevant work in the literature following the video approach is from Grauman et al. in [6, 11], where a summary methodology for egocentric video sequences is proposed. The authors rely on an initial event segmentation, followed by the detection of salient objects and people, create a graph linking events and the important objects/people, and finish with a selection of a subset of the events of interest. This final selection is based on combining three different measures: 1) *Story* (choosing a set of shots that are able to follow the inherent story in the dataset), 2) *Importance* (aimed at choosing only shots that show some important aspect of the day) and 3) *Diversity* (adding a way to avoid repeating similar actions or events in the summary). When considering the keyframe selection approach, one of the most relevant works is by Doherty et al. [5], where the authors study various selection methods like: 1) getting the frame in the middle of each segment, 2) getting the frame that is the most similar w.r.t. the rest of the frames in the event, or 3) selecting the closest frame to the event average.

3. METHODOLOGY

This section presents our methodology for keyframe-based summarization of egocentric photostreams, depicted in Figure 1. We start by characterizing each of the lifelog frames with a global scale visual descriptor. These features are used to create a visual-based event segmentation, which incorporates a post-processing step to guarantee time consistency. Finally, the most visually repetitive frame is selected as the most representative of the event.

3.1. Frames characterization

Convolutional Neural Networks (convnets or CNNs) have recently outperformed hand-crafted features in several computer vision tasks [12, 13]. These networks have the ability to learn sets of features optimised for a pattern recognition problem described by a large amount of visual data.

The last layer of these convnets is typically a soft-max classifier, which in some works is ignored, and the penultimate fully connected layers are directly used as feature vectors. These visual features have been successfully used as any other traditional hand-crafted features for purposes such as image retrieval [14] or classification [15].

In the field of egocentric video segmentation, convnets have also been proved as suitable for clustering purposes [10].

For this reason, we used a set of features extracted by means of the pre-trained *CaffeNet* convnet included in the Caffe library [16]. This convnet was inspired by [13] and trained on ImageNet [17]. In our case, we used as features the output of the penultimate layer, a fully connected layer of 4,096 components, discarding this way the final soft-max layer, which was intended to classify 1,000 different semantic classes from ImageNet.

3.2. Events segmentation

The egocentric photostream is segmented with an unsupervised hierarchical Agglomerative Clustering (AC) [18] based on the convnet visual features. As proved in [10], this clustering methodology reaches a reasonable accuracy for this task. In this way, we can define sets of images, each of them representing a different event. AC algorithms can be applied with different similarity measures. Different configurations were tested (see details in Section 4.2) and the best approach was obtained with the *average* linkage method with Euclidean distance. This option determines the two most similar clusters to be fused in each iteration using the following distance:

$$\arg \min_{C_i, C_j \in \mathbf{C}_t} D(C_i, C_j), \text{ where} \quad (1)$$

$$D(C_i, C_j) = \frac{1}{|C_i| \times |C_j|} \sum_{s_{k,i} \in C_i, s_{l,j} \in C_j} \sqrt{f(s_{k,i})^2 - f(s_{l,j})^2}, \quad (2)$$

where \mathbf{C}_t is the set of clusters at iteration t , $s_{k,i}$ and $s_{l,j}$ are the samples in cluster C_i and C_j , respectively, and $f(s)$ are the visual features extracted by means of the convnet.

However, creating the clusters based only on visual features often generates non-consistent solutions from a temporal perspective. Typically, images captured in the same scenario will be visually clustered as a single event despite corresponding to separate moments. For example, frames from the beginning of the day, (e.g. when the user takes the train for commuting to work) may be visually indistinguishable with other frames from the end of the day (e.g. when the user is going back home by train too). Additionally, another usual problem when relying only on visual features is that sometimes very small clusters can be generated, a result which should be avoided because an event is typically required to have a certain span in time (e.g. 3 minutes, in our work).

In order to solve these problems, we introduce two post-processing steps for refining the resulting clusters: *Division* and *Fusion*. The *Division* step splits in different events those images in the same cluster which are temporally interrupted by events defined in other clusters. For example, the event in orange from Figure 2 a) is divided in two events (orange and yellow) in Figure 2 c) due to a *Corridor scene* event (in green) interrupting the original *Office scene*. On the other hand, the second post-processing step, *Fusion*, will merge all those events shorter than a threshold with the closest neighboring event in time.

3.3. Keyframe selection

Once the photostream is split into the events, the next step is to carefully select a good subset of keyframes. To do so, we explored two different methods: *random walk* and a *minimum distance* approach. Both approaches are based on the assumption that the best photo to represent the event is the one, which is the most visually similar with the rest of the photos in the same cluster. As a result, each event can be automatically represented by a single image and, when all images combined, they will provide a visual summary of the user's day.

3.3.1. Random Walk

We propose to use the Random Walk algorithm [4] in each of the events, separately. As a result, the algorithm will select the photo, which is more visually similar to the rest of the photos in the event. After applying the same procedure for all the events, we can have a good general representation of the main events that happened in the user's daily life.

The Random Walk algorithm works as follows: 1) the visual similarity for each pair of photos in the event is computed; 2) a graph described by a transitional probabilities matrix is built using the extracted similarities as weights on each of the edges; 3) the matrix eigenvectors are obtained, and 4) the image associated to the largest value in the first eigenvector is considered as the keyframe of the event.

3.3.2. Minimum distance

The second considered option selects the individual frame with the minimal accumulated distance with respect to all the other images in the same event. That is, let us consider the adjacency matrix $A = \{a_{i,j}\} = \{d_{s_i,s_j}\}$, where d_{s_i,s_j} is the Euclidean distance between the descriptors of images s_i and s_j extracted by the convnet, $i = 1, \dots, N$, $j = 1, \dots, N$, where N is the number of frames of the event. Let us consider the vector $v = (\sum_j a_{i,j})$ of accumulated distances. One can easily see that the index of the minimal component of vector v i.e. $k = \arg \min_i \{v_i\}$, $i = 1, \dots, N$ determines the closest frame to the rest of frames in the corresponding event with respect to the L_1 norm [5].

4. RESULTS

This section presents the quantitative and qualitative experiments run on a home-made egocentric dataset to assess the performance of the presented technique.

4.1. Dataset

Our experiments were performed on a home-made dataset of images acquired with a Narrative¹ wearable camera. This device is typically clipped on the users' clothes under the neck

¹www.getnarrative.com



Fig. 2. Example of the events labeling produced by a) simply using the AC algorithm, b) applying the division strategy and c) additionally applying the fusion strategy. Each color represents a different event.

or around the chest area. The dataset, we used, is a subset of the one used by the authors in [10] (not using the SenseCam sets). It is composed of 5 day lifelogs of 3 different persons and has a total of 4,005 images. Furthermore, it includes the ground truth (GT) events segmentation for assessing the clustering results.

4.2. Quantitative evaluation of event segmentation

The first test assessed the quality of the photostream segmentation into events. In order to make this evaluation, we used the Jaccard Index, which is intended to measure the overlap of each of the resulting events and the GT the following way:

$$J(E, GT) = \sum_{e_i \in E, g_j \in GT} M_{ij} \frac{e_i \cap g_j}{e_i \cup g_j}, \quad (3)$$

where E is the resulting set of events, GT is the ground truth, e_i and g_j are a single event and a single GT segment respectively, and M_{ij} is an indicator matrix with values 1, iff e_i has the highest match with g_j .

We compared different cluster distance methods with respect to the chosen cut-off parameter (which determines how many clusters are formed considering their distance value) for the AC (see Figure 3). We choose the "average" with cutoff = 1.154 as the best option and, with this configuration, we measured the gain of introducing the Division-Fusion strategy, illustrated in Figure 4.

4.3. Qualitative evaluation with blind-test taste

The assessment of visual summaries of a day, like the one shown in Figure 5, is a challenging problem, because there is not a single solution for it. Different summaries of the same day may be considered equally satisfactory due to near duplicate images and subjectivity in the judgments. Therefore, we followed an evaluation procedure similar to the one adopted by Lu and Grauman [6]. We designed a blind-taste test and asked to a group of 20 people to rate the output of different solutions, without knowing which of them corresponded to each configuration.

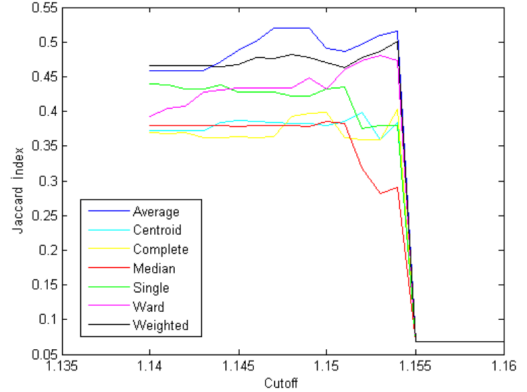


Fig. 3. Average Jaccard index value obtained for the 5 sets. We compare each of the methods after applying the division-fusion strategy with respect to the best cut-off AC values.

4.3.1. Keyframe selection

The first qualitative evaluation focused on the keyframe selection strategy, comparing both presented algorithms (*Random Walk* and *Minimum Distance*) with a third one, *Random Baseline*. In this first part, the three selection strategies were applied on each of the events defined by the GT annotation.

On the first part, we showed to the user a complete event according to the GT labels and, afterwards, the three keyframes selected by the three methods under comparison in a random sorting². Then, the user had to answer if each of the candidates was representative of the current event (results in Figure 6), and also choose which of them was the best one (results in Figure 7). This procedure was applied on each of the events of the day and results averaged per day.

Scoring results presented in Figures 6 and 7 indicate how both proposed solutions consistently outperform the random baseline for each day. The difference is more remarkable, when we asked the user to choose between only one of the possibilities (Figure 7). We must note that usually the result was very similar either for the *Random Walk* and the *Minimum Distance*, since in most of the cases both algorithms selected the same keyframe.

²If any of the results for the different methods was repeated, only one image was shown and the results were counted for both methods.

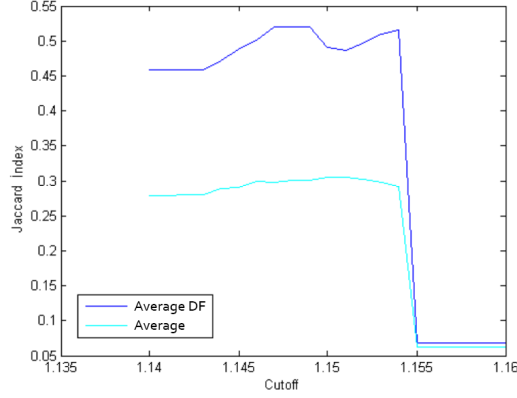


Fig. 4. Effect when using (dark blue) the division-fusion (DF) strategy and when not using it (light blue) in the average Jaccard index result for all the sets.

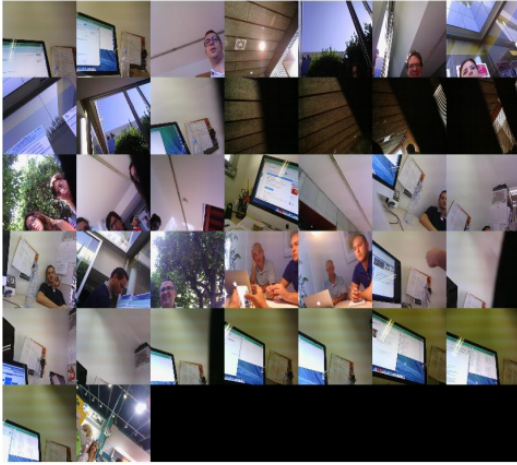


Fig. 5. Example of one of the summaries obtained by applying our approach on a dataset captured with Narrative camera.

4.3.2. Visual daily summary

In the second part of our qualitative study, we assessed the whole daily summary, built with the automatic event segmentation and the different solutions for keyframe selection. In this experiment, we added a fourth configuration that built a visual summary with a temporal *Uniform Sampling* of the day photostream, in such a way that the total amount of frames was the same as the amount of events detected through AC.

This time the user was shown the four summaries of the day generated by the four configurations. Figure 5 provides an example built with the *Random Walk* solution. For each summary, the user was firstly asked whether the set could represent the day (results in Figure 8), and also which of the four was the one that better described the day (results in Figure 9).

Focusing on the average results in Figure 5, we can state that, either applying *Random Walk* (88%) or *Minimal Distance* (86%), most of the generated summaries were posi-

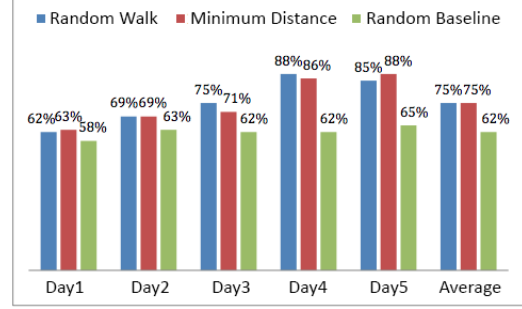


Fig. 6. Results answering "yes" to the question "Is this image representative for the current event?"

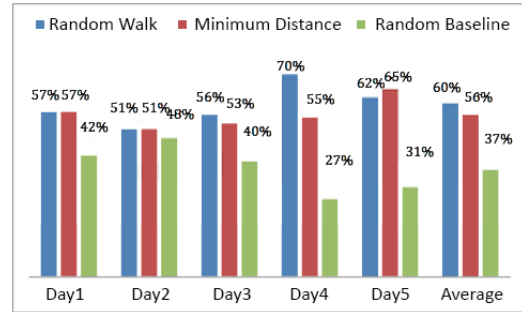


Fig. 7. Results to the question "Which of the previous frames is the most representative for the event?"

tively assessed by the graders. Moreover, when it comes to choose only the best summary, our method gathered 58% of the total votes if we consider that the voting is exclusive and that the summaries produced by the Random Walk and the Minima Distance methods are very similar. As a result, we obtained 34% and 41% of improvement respectively w.r.t the Random and the Uniform baselines.

5. CONCLUSIONS

In this work, we presented a new methodology to extract a keyframe-based summary from egocentric photostreams. After the qualitative validation made by 20 different users, we can state that our method achieves very good and representative summary results from the final user point of view.

Additionally, and always considering that the ultimate goal of this project is to reactivate the memory pathways of MCI patients, it offers satisfactory results in terms of capturing the main events of the daily life of the wearable camera users. A public-domain code developed for our visual summary methodology, is published in ³

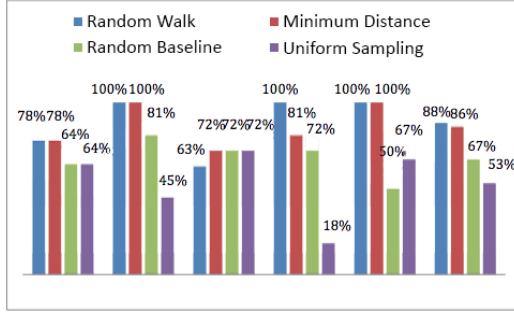


Fig. 8. Results answering “yes” to the question “Can this set of images represent the day?”

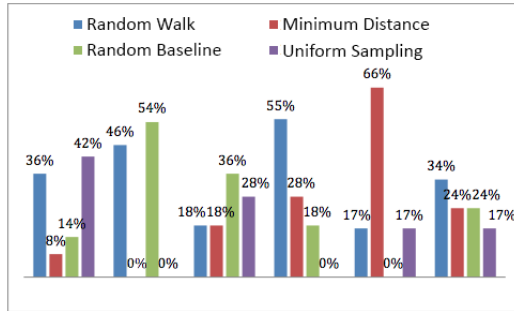


Fig. 9. Results to the question “Which of the previous summaries does better describe the day?”

6. REFERENCES

- [1] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood, “Sensecam: A retrospective memory aid,” in *UbiComp 2006: Ubiquitous Computing*, pp. 177–193. Springer, 2006.
- [2] Matthew L Lee and Anind K Dey, “Providing good memory cues for people with episodic memory impairment,” in *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 2007, pp. 131–138.
- [3] Wei-Hao Lin and Alexander Hauptmann, “Structuring continuous video recordings of everyday life using time-constrained clustering,” in *Electronic Imaging 2006*. International Society for Optics and Photonics, 2006, pp. 60730D–60730D.
- [4] Karl Pearson, “The problem of the random walk,” *Nature*, vol. 72, no. 1865, pp. 294, 1905.
- [5] Aiden R Doherty, Daragh Byrne, Alan F Smeaton, Gareth JF Jones, and Mark Hughes, “Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs,” in *Proceedings of the 2008 international conference on Content-based image and video retrieval*. ACM, 2008, pp. 259–268.
- [6] Z. Lu and K. Grauman, “Story-driven summarization for egocentric video,” in *CVPR*. 2013, pp. 2714–2721, IEEE.
- [7] M. Bolaños, M. Garolera, and P. Radeva, “Video segmentation of life-logging videos,” in *Articulated Motion and Deformable Objects*, pp. 1–9. Springer-Verlag, 2014.
- [8] Y. Poleg, Ch. Arora, and Shm. Peleg, “Temporal segmentation of egocentric videos,” in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference On*, 2014.
- [9] A. R. Doherty and A. F. Smeaton, “Automatically segmenting lifelog data into events,” in *Proceedings*, Washington, USA, 2008, WIAMIS ’08, pp. 20–23, IEEE Comp. Society.
- [10] E. Talavera, M. Dimiccoli, M. Bolaños, M. Aghaei, and P. Radeva, “R-clustering for egocentric video segmentation,” in *Iberian Conference on Pattern Recognition and Image Analysis (in press)*. Springer, 2015.
- [11] Y. J. Lee, J. Ghosh, and K. Grauman, “Discovering important people and objects for egocentric video summarization,” in *CVPR*. 2012, pp. 1346–1353, IEEE.
- [12] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS 25*, pp. 1097–1105. Curran Associates, Inc., 2012.
- [14] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky, “Neural codes for image retrieval,” in *Computer Vision–ECCV 2014*, pp. 584–599. Springer, 2014.
- [15] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the devil in the details: Delving deep into convolutional nets,” in *British Machine Vision Conference*, 2014.
- [16] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.

- [17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.
- [18] William HE Day and Herbert Edelsbrunner, “Efficient algorithms for agglomerative hierarchical clustering methods,” *Journal of classification*, vol. 1, no. 1, pp. 7–24, 1984.